Journal of Artificial Intelligence and Technology, (Ahead of Print)
https://doi.org/10.37965/jait.2026.1272

ISTP
RESEARCH ARTICLE

Agentic Artificial Intelligence for Zero-Day Cyber Threat Detection: An Adaptive Reasoning Approach

Thavavel Vaiyapuri[1] and Karthiyayini Murugesan[2]

[1]College of Computer Engineering and Science, Prince Sattam bin Abdulaziz University, Al Kharj, Saudi Arabia
[2]Department of Computer Science and Engineering, Sethu Institute of Technology, Kariapatti, Tamil Nadu, India

(Received 18 March 2026; Revised 20 April 2026; Accepted 24 May 2026; Published online 21 June 2026)

Abstract: Zero-day cyber threats continue to be a significant challenge for securing digital infrastructure in future economies. As such, there is a need for defensive systems that are intelligent and adaptive and, most importantly, trustworthy. However, traditional intrusion detection systems (IDSs) are typically static and lack the capability to reason, adapt, and explain their decisions. To address these limitations, this paper proposes an agentic artificial intelligence (AI)-driven framework for zero-day cyber threat detection based on adaptive reasoning. The framework improves zero-day awareness through autonomous reasoning, adaptive decision-making processes, and interpretable behavioral analysis. The key contribution of this research is the integration of three components, namely epistemic uncertainty estimation, embedding-based structural deviation assessment, and adaptive thresholding. These components allow the framework to make self-regulated decisions beyond fixed-model inference. The effectiveness of the framework was assessed using two benchmark datasets, Network Security Laboratory – Knowledge Discovery and Data Mining (NSL-KDD) and Telemetry and Network Traffic for Internet of Things (ToN-IoT). The experimental results demonstrate strong detection performance, improved class separability, reduced false-alarm rates (FARs), and consistent predictive confidence. Cross-dataset evaluation further demonstrates good generalization. In addition, interpretability analysis confirms that the framework relies on meaningful traffic characteristics rather than spurious correlations. In addition, the lightweight model design has the potential to support energy-efficient deployment in resource-constrained edge devices. Overall, the results demonstrate that the proposed framework provides a strong trade-off between accuracy, generalization, and interpretability, making it a promising solution for zero-day cyber threat detection in secure future economies.

Keywords: Adaptive threshold regulation; agentic intelligence; cybersecurity; energy-efficient AI; intrusion detection; structural embedding; uncertainty analysis

I. INTRODUCTION

Zero-day cyber threat detection is critical for securing future economies, where digital infrastructures must remain resilient against rapidly evolving and previously unseen attacks [1,2]. Modern cyber threats often emerge for the first time in real-world environments and are therefore not represented in training data. However, most existing intrusion detection system (IDS) assume that training data sufficiently captures future attack behavior. This is not true in practice [3]. New attack strategies often create new traffic patterns that differ significantly from those encountered during training, leading to misclassification of unseen threats [4]. This problem is even worse in heterogeneous network environments where there are many devices and many communication protocols [5]. Under such conditions, high detection accuracy alone is insufficient and IDS solutions must also remain reliable and adaptive under dynamic network conditions.

There are many barriers to developing effective IDSs for detecting zero-day cyber threats. One primary barrier is the phenomenon of distribution shift [1]. This can cause IDS to incorrectly classify traffic patterns that occur on a network, regardless of the level of confidence exhibited by the IDS. Another limitation is fixed decision boundaries [6]. An example of a fixed decision boundary is a static threshold. After the IDS has been deployed, the static threshold does not change. Thus, if the traffic pattern changes, then either false alarms increase or the number of attacks missed increases. The lack of interpretability is another limitation of IDSs [7]. Since the IDSs make decisions based on traffic patterns without explaining why certain actions were taken, the decisions made by the IDS are difficult to understand or trust by analysts. These limitations illustrate the need for IDS frameworks that can adapt to changing conditions while producing reliable and explainable decisions.

Artificial intelligence (AI) has recently provided new paths for improving the capabilities of IDSs. Some of the recent advancements include the utilization of deep learning (DL) techniques, unsupervised models, neural reasoning, and online learning (OL) [8,9]. Despite the emergence of these advancements, there are still many gaps that remain. Although some IDSs can operate in real time, the vast majority of IDSs do not have the capability to continuously adapt or modify their decision thresholds in response to changing traffic patterns [10]. Additionally, the prediction uncertainty is usually not included in the decision-making process. Thus, the output of the IDS is likely to be overconfident and unreliable [11,12]. Lastly, the explanation of the IDS is usually an afterthought instead of being integrated into the decision-making pipeline [13,14]. These gaps show that there is a need for intelligent IDS frameworks that can reason, adapt, and give understandable insights, especially in zero-day scenarios.

Corresponding author: Thavavel Vaiyapuri (e-mail: t.thangam@psau.edu.sa).

To address these challenges, this paper proposes an agentic AI-driven framework for zero-day cyber threat detection based on an adaptive reasoning approach, aligned with the requirements of secure future economies. The proposed framework introduces three key capabilities. First, it estimates epistemic uncertainty to quantify the reliability of predictions. Second, it uses embedding-based structural deviation analysis to find deviations that could represent previously unseen attack patterns. Third, it includes an adaptive thresholding mechanism that adjusts the decision boundary according to the changes in traffic. Jointly, these three capabilities enable the proposed framework to make self-regulated and adaptive decision-making beyond the typical fixed-model inference while also providing interpretable output.

The performance of the proposed framework was evaluated on the NSL-KDD and ToN-IoT benchmark datasets. The evaluation demonstrated the superiority of the proposed framework relative to state-of-the-art approaches in terms of detection accuracy, discriminative capability, false alarm rate (FAR), and predictive confidence. Additionally, cross-dataset evaluations demonstrated the generalization capability of the proposed framework. Finally, the interpretability analyses demonstrated that the proposed framework relied on meaningful traffic characteristics, rather than spurious correlations.

The main contributions of this work are summarized as follows:

a. An agentic AI-driven IDS framework for zero-day cyber threat detection that supports adaptive decision-making and interpretable outputs.

b. Integration of epistemic uncertainty estimation and embedding-based structural deviation analysis to improve reliability when encountering unseen attacks.

c. An adaptive thresholding mechanism that overcomes fixed decision boundaries and enhances decision stability under dynamic network conditions.

d. A comprehensive evaluation using NSL-KDD and ToN-IoT datasets with uncertainty and interpretability analysis.

II. LITERATURE REVIEW

Zero-day intrusion detection has progressed from simple threshold-based decisions to more adaptive detection mechanisms. Early studies classified traffic as zero-day when anomaly scores or behavioral deviations exceeded predefined or optimized thresholds. For example, [15] optimized the anomaly-score threshold, [16] learned evolving temporal behavioral deviations, and [17] introduced a two-stage cascade threshold to first flag suspicious traffic and then reduce false positives. Thus, [15–17] show a clear shift from single-threshold detection toward adaptive and cascade-based threshold regulation.

However, threshold-based detection alone is not sufficient for dynamic zero-day environments. To address changing traffic behavior, [18] introduces adaptive self-adjusting memory to update the decision boundary under gradual, recurring, and incremental drift. This improves adaptability, but sudden drift remains difficult to handle. Beyond adaptive boundaries, [19] shifts the focus toward structural representation learning by using contrastive loss to improve separation between benign, known attack, and unseen zero-day traffic. Thus, [18,19] move the field from threshold selection toward adaptive and embedding-based generalization, but they still mainly produce prediction outputs without deeper explanation or analyst-oriented reasoning.

More recent studies have attempted to move from prediction models toward agent-based and agentic IDS. In [10], a deep reinforcement learning (DRL)-based NIDS with stacked long short-term memory (LSTM) is used to detect unseen attacks through adaptive learning. However, it mainly functions as a predictive model rather than a coordinated reasoning system. In contrast, [20] moves toward agentic AI by integrating uncertainty estimation, structural embedding analysis, and human-in-the-loop decision support for zero-day detection.

Table I summarizes the main approaches, contributions, and limitations of the most relevant studies. As shown in Table I, existing methods usually focus on one aspect, such as anomaly-score thresholding, drift-aware adaptation, adaptive learning, latent representation, or human-in-the-loop reasoning. Therefore, fragmented decision-evidence analysis remains a key limitation in zero-day IDS. This motivates the proposed agentic-driven adaptive reasoning framework, which integrates threat score, epistemic uncertainty, structural deviation, and adaptive threshold regulation into a unified decision pipeline.

III. PROPOSED FRAMEWORK

Zero-day cyberattack detection is achieved through an agentic AI architecture (Fig. 1), which integrates three major functionalities: perception, reasoning, and adaptive control. This integration enables modeling network behavior over time, analyzing deviation from normal traffic behavior, and adapting decisions as network conditions change. The goal is to maintain reliable detection even when traffic patterns evolve.

A. PERCEPTION AGENT

The perception agent captures network traffic as a temporal sequence rather than as an independent record. The i^{th} traffic

Table I. Summary of research gaps in existing zero-day IDS studies

Ref.	Approach	Key contribution	Limitation
[15–17]	Unsupervised learning	Optimized anomaly thresholds	Focus mainly on threat score
[18]	Self-adaptive kNN	Drift-aware adaptation	Limited reasoning and sensitive to neighbor selection
[10]	LSTM + DRL	DRL-based adaptive learning	Limited confidence analysis
[19]	Contrastive learning	Improved latent separation	Focus mainly on latent representation learning
[20]	Agentic visual analytics	Human-in-the-loop reasoning	Depends on analyst involvement
Our study	Agentic-driven adaptive reasoning	Integrates threat score, epistemic uncertainty, structural deviation, and adaptive threshold regulation	Addresses fragmented decision-evidence analysis in zero-day IDS

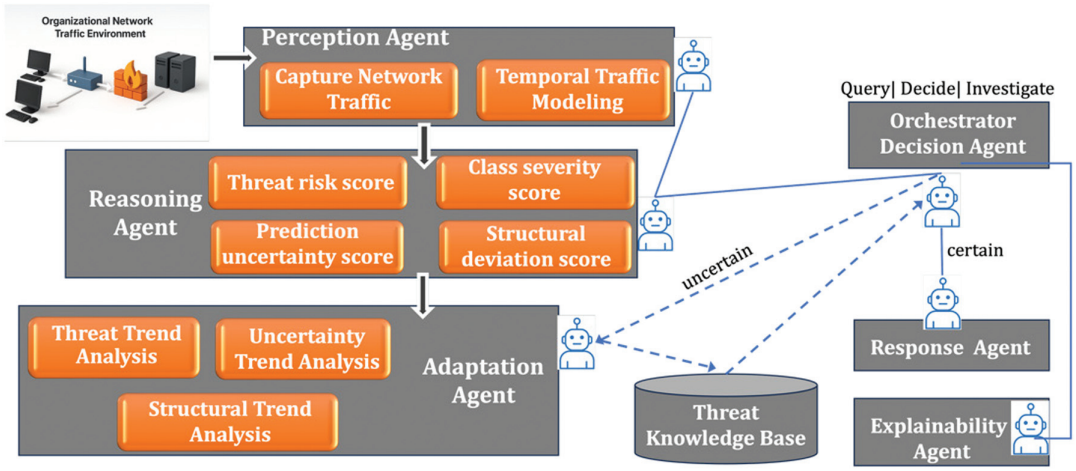


Fig. 1. Class distribution of the dataset used in this study.

sequence containing T consecutive vectors is represented as $X^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)})$where each vector is obtained from flow-level or connection-level data [21]. This sequential representation helps capture short-term behavioral changes that may indicate evolving or unknown attack activity. Temporal dependencies are encoded using an LSTM network. At each time step, the hidden state is updated as [8] and [12]:

$$h_t^{(i)} = LSTM(x_t^{(i)}, h_{t-1}^{(i)}) \tag{1}$$

where $x_t^{(i)}$ denotes the current traffic vector and $h_{t-1}^{(i)}$ denotes the previous hidden state. The updated state $h_t^{(i)}$ summarizes the traffic behavior observed up to time t. After processing the full sequence, the final hidden state $h_T^{(i)}$ is projected through a softmax layer:

$$\hat{p}^{(i)} = Softmax(Wh_T^{(i)} + b) \tag{2}$$

where W and b denote the trainable weight matrix and bias term, respectively. The output $\hat{p}^{(i)}$ provides a probability distribution over predefined traffic classes C. This output is passed to the reasoning agent for prediction, uncertainty estimation, and deviation scoring, enabling the framework to identify known attacks and detect subtle behavioral shifts associated with zero-day threats.

B. REASONING AGENT

The reasoning agent receives the probability output $\hat{p}^{(i)}$ and latent vector $h_T^{(i)}$ from the perception agent for each traffic sequence $X^{(i)}$. Instead of relying solely on the predicted class label, it derives evidence signals that quantify threat level, prediction reliability, attack severity, and structural abnormality [22]. These signals are then routed to the orchestrator decision agent for final decision-making.

1). THREAT RISK SCORE. The threat risk score quantifies how strongly an observed traffic sequence deviates from legitimate network behavior. It is derived from the probability assigned to the normal class by the perception agent [23]. For a given posterior distribution produced by the perception layer, the threat score (TS) is defined as:

$$TS^{(i)} = 1 - \hat{p}_N^{(i)} \tag{3}$$

where $\hat{p}_N^{(i)}$ denotes the predicted probability that $X^{(i)}$ belongs to the normal traffic class. When this probability is high, the traffic is considered consistent with benign behavior, resulting in a low threat score. Conversely, a lower normal-class probability produces a higher $TS^{(i)}$, indicating stronger evidence of anomalous or malicious activity.

2). PREDICTION UNCERTAINTY SCORE. The prediction reliability of the model is estimated using epistemic uncertainty through Monte-Carlo dropout. During inference, dropout remains active and the model performs K stochastic forward passes for the same traffic sequence $X^{(i)}$. This produces a set of posterior outputs: $\mathcal{P}^{(i)} = \{\hat{p}^{(i,1)}, \hat{p}^{(i,2)}, \ldots, \hat{p}^{(i,K)}\}$. The predictive mean distribution is computed as [11,12]:

$$\bar{p}^{(i)} = \frac{1}{K}\sum_{k=1}^{K} \hat{p}^{(i,k)} \tag{4}$$

Epistemic uncertainty U_{MC} is then quantified using the variance of these predictions which is computed as:

$$U_{MC}^{(i)} = \frac{1}{K}\sum_{k=1}^{K} \|\hat{p}^{(i,k)} - \hat{p}^{(i)}\|^2 \tag{5}$$

A higher variance indicates a larger amount of uncertainty in the model's prediction. In general, this is true when the input pattern does not match those patterns learned during training, which is commonly related to emerging or new attack patterns.

3). STRUCTURAL DEVIATION SCORE. Structural deviation score S_{emb} evaluates whether a traffic sequence is consistent with the learned latent-space structure. For each sequence $X^{(i)}$, the final hidden representation $h_T^{(i)}$ is used as its behavioral embedding. The distance between two traffic sequences $X^{(i)}$ and $X^{(j)}$ is computed as [24]:

$$D(X^{(i)}, X^{(j)}) = \|h_T^{(i)} - h_T^{(j)}\| \tag{6}$$

To measure structural deviation, the average distance from $X^{(i)}$ to its k-nearest neighbors (KNNs) in latent space is defined as [25]:

$$S_{emb}^{(i)} = \frac{1}{k}\sum_{X^{(j)} \in N_k(X^{(i)})} D(X^{(i)}, X^{(j)}) \tag{7}$$

Here, $\mathcal{N}_k(X^{(i)})$ denotes kNN set of $X^{(i)}$. A higher $S_{emb}^{(i)}$ indicates that the instance lies in a sparse or weakly supported region of the embedding space. Conversely, a lower value

corresponds to compact and well-structured neighborhoods that are consistent with known behavioral patterns. This structural reasoning complements the threat score and epistemic uncertainty measures by providing an additional and independent source of evidence regarding behavioral conformity or deviation in a traffic instance.

C. ORCHESTRATION DECISION AGENT

The orchestrator decision agent controls the final decision flow for each traffic sequence $X^{(i)}$. It receives the scores from the reasoning agent and evaluates the zero-day decision rule:

$$TS^{(i)} > \tau_{TS} \wedge U_{MC}^{(i)} > \tau_U \wedge S_{emb}^{(i)} > \tau_{emb} \tag{8}$$

Here, τ_{TS}, τ_U, and τ_{emb} denote calibrated thresholds for the threat score, epistemic uncertainty, and embedding-based structural deviation, respectively. Sequences satisfying all three conditions are treated as zero-day attack and are forwarded to the response agent. If Eq. (8) is not satisfied, the orchestrator checks the uncertainty level. When uncertainty is low, the predicted known-class label is retained and forwarded to the response agent. When uncertainty is high, the orchestrator activates the adaptation agent for threshold refinement. The updated thresholds are then used to re-evaluate $X^{(i)}$, as represented by the dotted feedback lines in Fig. 1. At the same time, the corresponding values of $TS^{(i)}$, $U_{MC}^{(i)}$, and $S_{\text{emb}}^{(i)}$ are stored in sliding windows for subsequent trend monitoring and adaptive threshold updating.

D. ADAPTATION AGENT

The adaptation agent updates the zero-day decision thresholds as network behavior changes over time. As shown in Fig. 1, it performs threat trend analysis, uncertainty trend analysis, and structural trend analysis using the recent score values generated by the reasoning agent. These values are stored in sliding windows and used to update the corresponding thresholds online [26]. At update step t, $W_{TS}^{(t)}$, $W_U^{(t)}$, and $W_{emb}^{(t)}$ represent the sliding-window sets containing recent values of $TS^{(i)}$, $U_{MC}^{(i)}$, and $S_{emb}^{(i)}$, respectively. The thresholds are adapted using an exponentially smoothed percentile-based rule. The threat score threshold is updated as [6]:

$$\tau_{TS}^{(t+1)} = (1-\eta)\tau_{TS}^{(t)} + \eta Q_\alpha(\mathcal{W}_{TS}^{(t)}) \tag{9}$$

and the prediction uncertainty threshold is updated as:

$$\tau_U^{(t+1)} = (1-\eta)\tau_U^{(t)} + \eta Q_\alpha(\mathcal{W}_U^{(t)}) \tag{10}$$

In the same way, the embedding-based structural threshold is modified as [27]:

$$\tau_{emb}^{(t+1)} = (1-\eta)\tau_{emb}^{(t)} + \eta Q_\alpha(W_{emb}^{(t)}) \tag{11}$$

where $Q_\alpha(\cdot)$ denotes the empirical α-percentile of the corresponding sliding window and $\eta \in (0,1)$ controls the adaptation rate. This mechanism allows the rejection boundaries to evolve with changes in benign traffic behavior, attack activity, and latent-space structure. As a result, outdated thresholds are avoided, robustness to behavioral drift is improved, and reliable zero-day detection is maintained during long-term deployment. The updated thresholds and trend summaries are then forwarded to the orchestrator decision agent and threat knowledge base for subsequent decision support. The overall workflow of proposed framework is described in Algorithm 1.

Algorithm 1. Agentic AI for early zero-day cyberattack detection

Input: Network traffic stream $\{X^{(i)}\}_{i=1}^{N}$; initial thresholds $\tau_{TS}^{(0)}$, $\tau_U^{(0)}$, $\tau_{emb}^{(0)}$; adaptation rate η; sliding-window size m.

Output: Detection label for each traffic instance

For each incoming traffic sequence $X^{(i)}$ *do*

Perception Agent: compute $h_T(X^{(i)})$ and posterior output $\hat{p}(X^{(i)})$.

Reasoning Agent: compute $TS^{(i)}, U_{MC}^{(i)}, S_{emb}^{(i)}$.
Append $\{TS^{(i)}, U_{MC}^{(i)}, S_{emb}^{(i)}\}$ to $\{W_{TS}, W_U, W_{emb}\}$

Decision Agent:

i. Apply the zero-day decision rule defined in Eq. (8).

ii. If Eq. (8) is satisfied, X_i is zero-day;

iii. If Eq. (8) is not satisfied and $U_{MC}^{(i)}$ is low, retain the predicted label.

Else mark $X^{(i)}$ as uncertain

Adaptation Agent: When activated, use the current sliding-window statistics to update τ_{TS}, τ_U, τ_{emb} using Eq. (9)–(11).

Response Agent: Generate the final security response and provide an interpretable explanation for the final decision

E. RESPONSE AGENT

The response agent executes the final operational action after the orchestrator decision agent confirms the decision for traffic sequence $X^{(i)}$. Based on the predicted threat type and severity level, it can trigger actions such as alert generation, traffic blocking, connection isolation, logging, or priority escalation. This enables rapid mitigation of confirmed malicious activity while preserving normal traffic continuity.

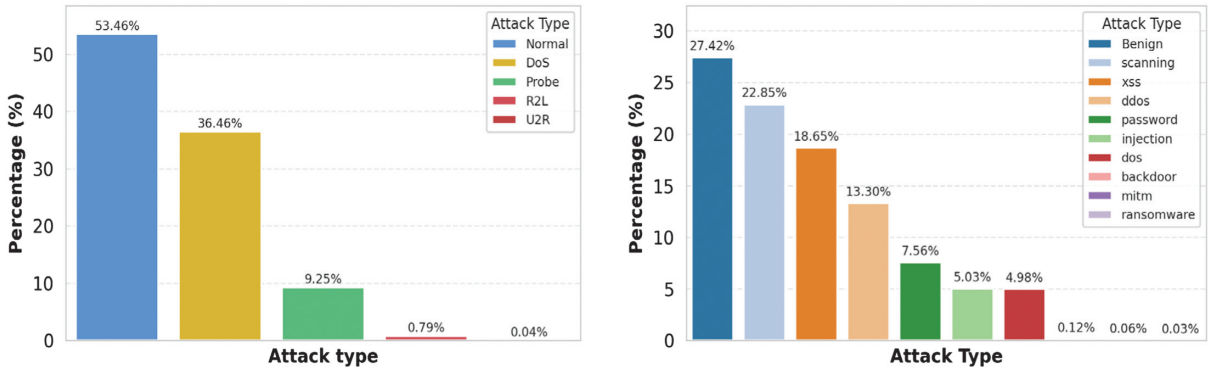
The explainability agent provides human-interpretable justification for the final decision. It summarizes the main factors that influenced the detection of $X^{(i)}$, including the dominant risk evidence and confidence level derived from the preceding agents. This information supports analyst verification, improves transparency, and assists forensic investigation.

IV. EXPERIMENTAL SETUP

A. DATASET DESCRIPTION

This paper examines the performance of the proposed intrusion detection framework on two benchmark datasets: NSL-KDD [28] and ToN-IoT [29]. These datasets were selected because they represent different network environments and attack behaviors, thereby enabling a comprehensive assessment of the detection capability of the proposed model. The NSL-KDD dataset has many improvements over the original KDD'99 dataset. One of these improvements includes removing redundant traffic and reducing learning biases in the traffic data. The dataset includes normal and malicious network traffic where the level of complexity for each type of attack varies. In this evaluation, we have chosen to utilize the class distribution illustrated in Fig. 2(a).

The ToN-IoT dataset is based on large-scale traffic captured from Internet of Things (IoT) environments. It includes traffic from heterogeneous data sources and diverse cyberattack types. The ToN-IoT dataset is more complex and also has a much larger imbalance in its class distribution compared to the NSL-KDD dataset. The class distributions for the ToN-IoT dataset used in this research are presented in Fig. 2(b). Use of these two datasets help compare the model performance performs in a traditional network



(a) NSL-KDD (b) ToN-IoT

Fig. 2. Class distribution of the dataset used in this study.

environment as well as in modern IoT environment. Thus, these datasets provide an opportunity to assess the model robustness and generalizability across different types of network environments.

B. ZERO-DAY EVALUATION PROTOCOL

An open-set evaluation protocol was used to assess whether the proposed framework can detect attacks absent during training. The model was trained only on benign traffic and selected known attack classes, while specific attack categories were completely excluded from both training and validation and introduced only during testing. The excluded classes were treated as zero-day attacks. They were not used as supervised output classes during model training. Instead, they were identified at inference time by the orchestrator decision agent using the combined evidence of threat score, epistemic uncertainty, and structural deviation. For NSL-KDD, benign and Denial of Service (DoS) traffic were used for training, while Probe, R2L, and U2R were held out as unseen zero-day attacks. For ToN-IoT, backdoor, Man-in-the-Middle Attack (MITM), and ransomware were held out for zero-day testing.

C. LSTM ARCHITECTURE

The perception module uses lightweight bidirectional LSTM (BiLSTM) network to extract the short-term temporal characteristics of the network traffic [21]. This lightweight configuration was selected to reduce computational overhead and support efficient edge deployment. Each flow record is represented using a single-timestep feature vector with all input attributes (F). The stacked BiLSTM layers extract temporal dependencies in both forward and backward directions. After identifying these temporal relationships, the output of each layer is fed into a dense layer, and a softmax activation function is applied to map the output to one of the target classes. The network was trained utilizing Adam optimizer with categorical cross-entropy loss. A summary of the network architecture can be found in Table II.

Table II. LSTM architecture configuration

Layer	Configuration
Input	(1, F) where F is number of features
BiLSTM	3 stacked layers, each with 64 units
Dropout	Dropout = 0.2 between layers
Dense	64 neurons with ReLU
Output	Softmax classifier
Loss	Categorical cross-entropy
Optimizer	Adam with learning rate = 0.001
Training	Mini batch with early stopping

V. RESULTS AND DISCUSSION

A. ABLATION ANALYSIS

To evaluate the contribution of each component in the proposed framework, an ablation analysis was conducted on the NSL-KDD dataset. The baseline model was first evaluated, and then TS, U_{MC}, S_{emb}, and adaptive threshold were progressively added. Table III reports the class-wise precision (Pre), detection rate (DR), F1-score, and FAR for benign, DoS, Probe, and zero-day classes.

Table III. Ablation analysis results on NSL-KDD

Variants	Class	Pre	DR	F1	FAR
Baseline	Normal	0.87	0.98	0.92	0.138
	DoS	0.99	0.95	0.97	0.005
	Probe	0.89	0.91	0.90	0.011
	Zero-Day	0.0	0.0	0.0	0.0
Baseline + TS + U_{MC}	Normal	0.89	0.98	0.93	0.117
	DoS	0.99	0.95	0.97	0.005
	Probe	0.89	0.91	0.90	0.011
	Zero-Day	0.29	0.07	0.11	0.008
Baseline + TS + U_{MC} + S_{emb}	Normal	0.90	0.98	0.94	0.106
	DoS	0.99	0.95	0.97	0.004
	Probe	0.97	0.91	0.94	0.003
	Zero-Day	0.20	0.11	0.13	0.020
Proposed method	Normal	0.99	1.00	1.00	0.001
	DoS	1.00	1.00	1.00	0.000
	Probe	1.00	1.00	1.00	0.000
	Zero-Day	0.83	0.98	0.89	0.010

The baseline model performs well on known classes, achieving high F1-scores. However, it completely fails to detect zero-day attacks. This shows that the baseline closed-set classifier is unable to recognize unseen attack samples and instead misclassifies them into known categories.

After adding TS and U_{MC}, the model begins to detect zero-day samples, improving Pre, DR, and F1-score to 0.29, 0.07, and 0.11, respectively. This indicates that TS and U_{MC} help identify some unknown samples that the baseline model misses. However, the low DR shows that many zero-day samples are still not detected. When S_{emb} is added, the model further improves the structural deviation in the latent feature space.

When S_{emb} is added, the model further improves the representation of structural deviation in the latent feature space. Probe class improves notably, with Pre increasing to 0.97 and F1-score increasing to 0.94. The zero-day DR also increases from 0.07 to 0.11, and F1-score increases from 0.11 to 0.13. Although zero-day precision decreases from 0.29 to 0.20 and FAR increases to 0.020, this indicates that the model becomes more sensitive to unseen attacks, detecting more zero-day instances at the cost of additional false alarms.

The proposed method which includes adaptive threshold regulation achieves the best overall performance. It obtains near-perfect results for benign, DoS, and Probe classes with F1-scores of 1.00 and almost zero FAR. More importantly, zero-day detection improves substantially reaching 0.83 Pre, 0.98 DR, and 0.89 F1-score, with a low FAR of 0.010. These results show that adaptive threshold regulation is essential for balancing sensitivity and false alarms. Overall, the ablation study confirms that each component contributes to the final performance, while the complete proposed method provides the most reliable detection of both known and zero-day attacks.

B. UNCERTAINTY ANALYSIS

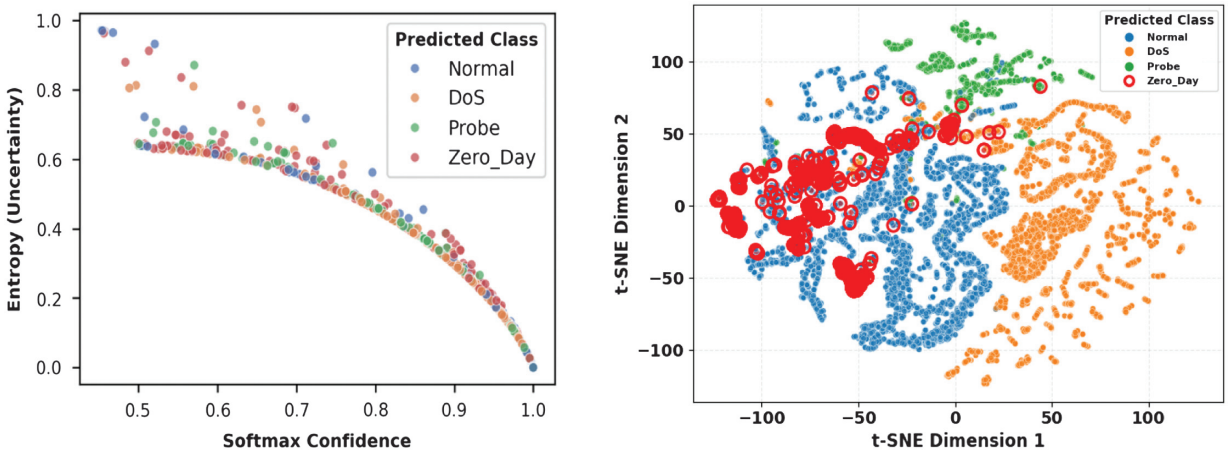
The uncertainty behavior of the model on the NSL-KDD dataset is illustrated in Fig. 3. The purpose of this analysis is to examine whether the model's confidence is consistent with its predictive uncertainty and whether uncertain samples are concentrated in meaningful regions of the feature space. Figure 3(a) shows the relationship between softmax confidence and entropy. A clear inverse pattern can be observed: samples with high softmax confidence are generally associated with low entropy, whereas samples with lower confidence show higher entropy. This indicates that the model produces more certain predictions when the posterior probability is concentrated in one class, while uncertainty increases when the prediction distribution becomes less decisive. The gradual decrease in entropy as confidence increases also suggests that the model's uncertainty behavior is stable rather than irregular.

Figure 3(b) visualizes the latent feature representation using t-distributed stochastic neighbor embedding (t-SNE). The visualization was generated by projecting the learned feature embeddings into a two-dimensional space using a perplexity value of 30 with Principal Component Analysis (PCA)-based initialization and a fixed random seed of 42 to ensure reproducibility. The resulting projection shows that normal, DoS, and Probe samples form relatively distinguishable clusters, whereas the highlighted zero-day samples are primarily concentrated near overlapping or boundary regions between known classes.

This behavior indicates that zero-day samples are structurally closer to ambiguous regions of the latent space rather than well-separated class centers. Overall, the results show that the model assigns lower uncertainty to confident known-class predictions and higher uncertainty to samples located near class-overlapping regions. These observations support the use of entropy and latent-space deviation as complementary indicators for regulating decisions and identifying potential zero-day intrusions.

C. CROSS-DATASET EVALUATION

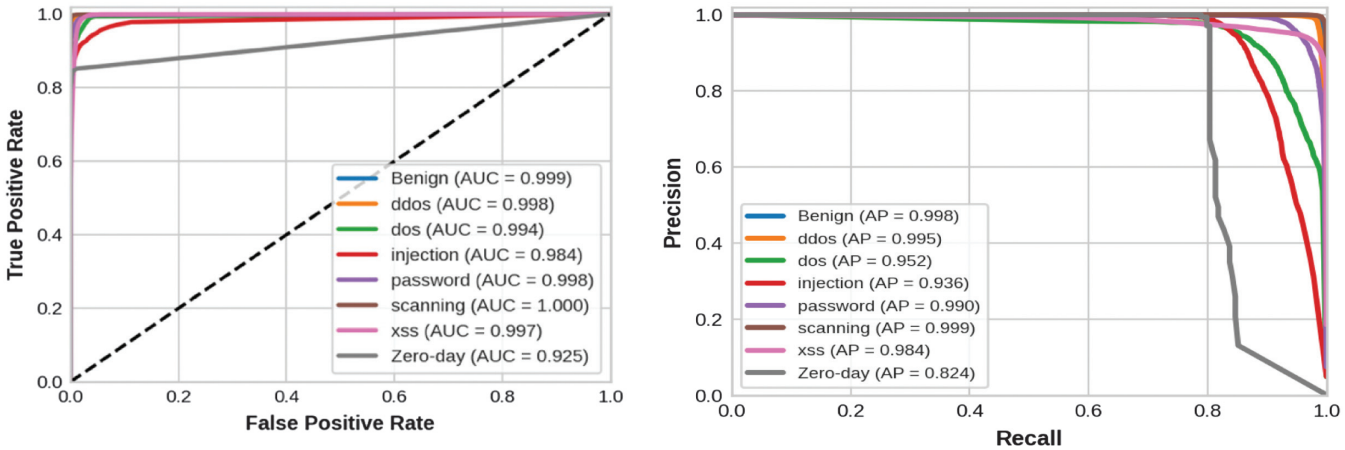
To evaluate whether the proposed framework generalizes beyond the training distribution, the model was tested on the ToN-IoT dataset, which contains heterogeneous traffic patterns, diverse attack behaviors, and additional zero-day attack categories. Receiver operating characteristic (ROC) and precision–recall (PR) analyses were employed to examine the discrimination capability and reliability of the framework under cross-dataset conditions. Figure 4(a) shows the ROC curves generated by varying the classification threshold and measuring the corresponding true positive and false positive rates for each class. The known classes such as Cross-Site Scripting (XSS) achieve area under the



(a) Confidence–entropy relationship

(b) t-SNE visualization

Fig. 3. Uncertainty analysis on NSL-KDD dataset.



(a) ROC Curve (b) PR Curve

Fig. 4. Cross-dataset discriminative analysis on the ToN-IoT dataset.

curve (AUC) values above 0.98, indicating strong separability. The zero-day class obtains an AUC of 0.925, showing that the framework can reasonably distinguish unseen attacks under cross-dataset conditions.

Figure 4(b) presents the PR curves, which evaluate the trade-off between attack detection capability and FAR across different decision thresholds. Most known classes maintain high precision over a broad recall range, with average precision (AP) values close to 1.0. DoS and injection achieve AP values of 0.952 and 0.936, respectively, while the zero-day class records an AP value of 0.824. The lower AP value for zero-day traffic reflects the increased difficulty of detecting unseen attacks under cross-dataset conditions, although the framework still maintains relatively high precision while identifying a substantial portion of unknown attacks.

Figure 5 further summarizes the reliability characteristics of the detection framework using a reliability quadrant plot based on recall and false positive rate values. Classes positioned in the reliable region exhibit both high recall and low false positive rates. Most known classes are concentrated in this region, indicating stable detection performance. The zero-day appear in the cautious region, suggesting moderate recall while maintaining low FAR.

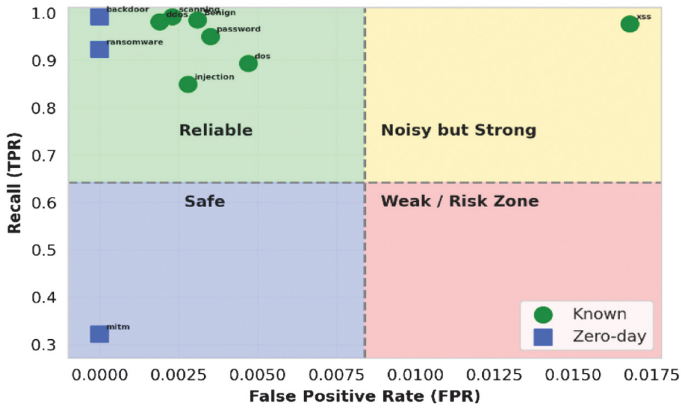


Fig. 5. Reliability quadrant analysis on ToN-IoT.

Table IV. Quantitative analysis on ToN-IoT dataset

Class	Pre	DR	F1	FAR
Benign	0.992	0.985	0.988	0.003
DDoS	0.987	0.982	0.984	0.002
DoS	0.909	0.893	0.901	0.005
Injection	0.941	0.849	0.893	0.003
Password	0.957	0.950	0.954	0.004
Scanning	0.990	0.992	0.991	0.003
XSS	0.930	0.977	0.953	0.017
Zero-Day	0.982	0.631	0.769	0.000

In contrast, the MITM category exhibits lower recall, indicating that some attacks remain undetected; however, its low false positive rate suggests that the framework behaves conservatively and avoids incorrectly classifying legitimate traffic as malicious. Consistent with the visual results, Table IV shows strong performance across most classes. Benign, Distributed DoS (DDoS), password, scanning, and XSS achieve F1-scores above 0.95, while DoS and injection obtain slightly lower F1-scores of 0.901 and 0.893, respectively. The zero-day class achieves high precision of 0.982 but lower DR of 0.631, resulting in an F1-score of 0.769. Its FAR is 0.000, indicating that the model detects zero-day attacks conservatively without incorrectly flagging normal or known-class samples as zero-day. In summary, the cross-dataset results show that the proposed framework remains effective under distribution shifts. It achieves strong detection performance for known attacks and provides a reliable indication of unknown or zero-day traffic. These findings suggest its potential suitability for real-world IoT network security.

D. INTERPRETABILITY ANALYSIS

Interpretability was assessed not only to visualize the model decision but also to examine whether the explanation remains reliable under minor input changes. This is important in trustworthy AI, where explanations should be consistent and useful for analyst verification in security-critical systems [30]. Therefore,

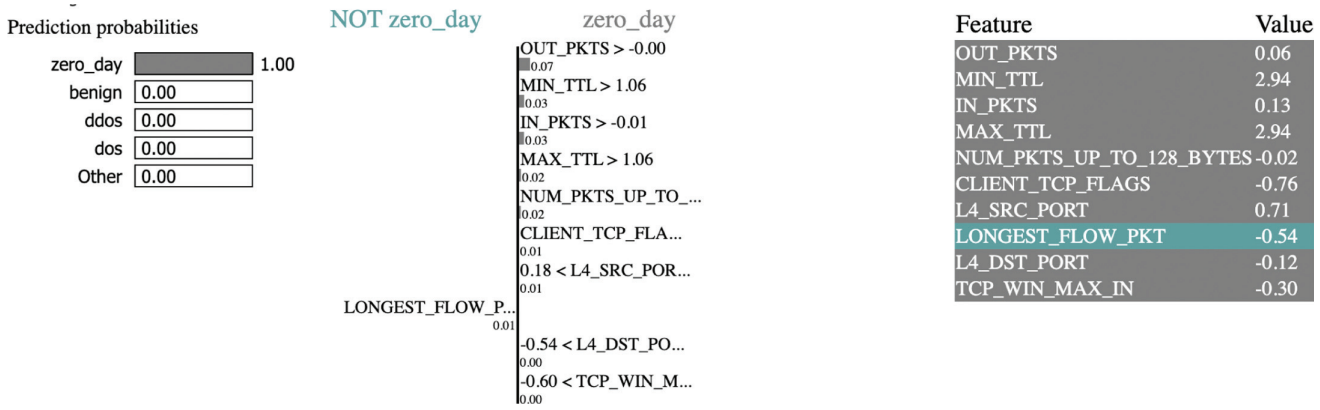


Fig. 6. Interpretability analysis of zero-day decision behavior before input perturbation.

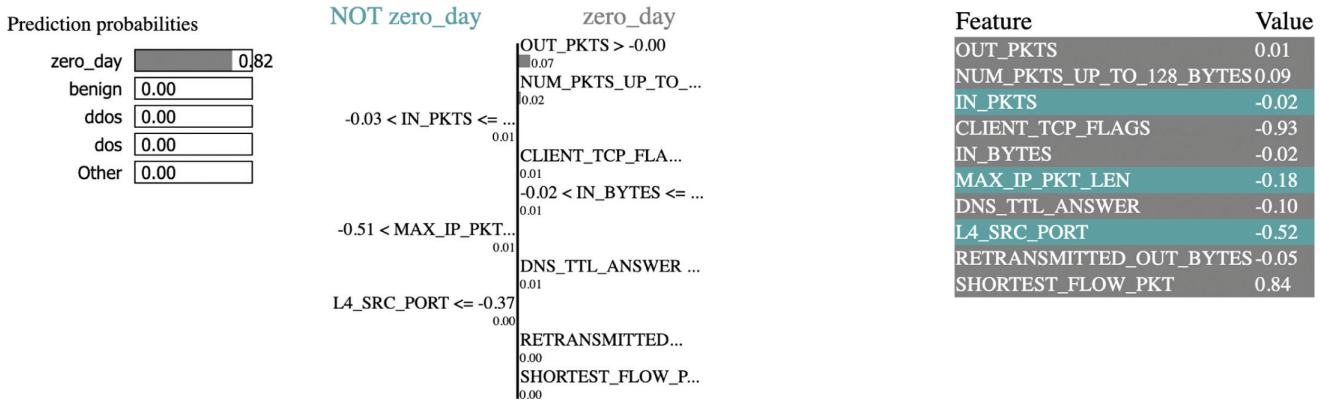


Fig. 7. Interpretability analysis of zero-day decision behavior after input perturbation.

Figs. 6 and 7 present the explanation generated for a zero-day sample before and after a small input perturbation, respectively.

In both cases, the model maintained the zero-day prediction with high confidence, indicating stable decision behavior under minor feature variations. Although the prediction confidence decreased slightly after perturbation, the dominant contributing features remained largely consistent across both explanations.

Features related to packet statistics, flow characteristics, Transmission Control Protocol (TCP) flags, and port behavior, such as OUT_PKTS, IN_PKTS, CLIENT_TCP_FLAGS, and L4_SRC_PORT, continued to contribute significantly to the zero-day decision. The consistency of these explanations suggests that the framework does not rely on isolated noisy features but instead captures structurally meaningful traffic characteristics associated with abnormal behavior. This observation provides preliminary evidence that the interpretability behavior remains reasonably stable under small perturbations, rather than producing completely inconsistent explanations for similar inputs.

From an operational perspective, these explanations can support a human-in-the-loop workflow by providing analysts with interpretable evidence regarding why a traffic instance was flagged as zero-day. The feature attribution maps help identify which traffic characteristics contributed most strongly to the detection decision, thereby improving transparency and supporting analyst verification in security-sensitive IoT environments.

Overall, the interpretability analysis indicates that the framework not only produces accurate predictions but also generates relatively consistent and explainable decision patterns that may improve trust and usability in practical intrusion detection scenarios.

E. COMPARISON WITH RELATED WORKS

An effective evaluation of a zero-day IDS should consider both classification performance and generalization under unseen traffic conditions. A dependable system must detect attack patterns that differ from those observed during training. Several studies have explored adaptive intrusion detection approaches including hybrid OL, DRL, unsupervised autoencoders, and neural reasoning methods.

Although these methods perform well on benchmark datasets, most do not update their model behavior or decision thresholds after training. Therefore, their detection reliability may decrease when new attack patterns appear.

Table V compares the proposed framework with related intrusion detection methods across benchmark datasets. The proposed method achieves 97.4% accuracy on ToN-IoT and 98% accuracy on NSL-KDD, which is comparable to strong existing methods such as hybrid OL [31], LSTM with DRL [10], self-adaptive kNN [32], quantum support vector machine (QSVM)

Table V. Comparison with related works

Ref.	Approach	Dataset	Acc
[31]	Hybrid OL	IBM	98.4%
		NSL-KDD	96.6%
[10]	LSTM + DRL	ToN-IoT	99%
		UNSW-NB15	95%
		BoT-IoT	97%
[32]	Unsupervised autoencoder	CICIDS2017	75–98%
		NSL-KDD	89–99%
[18]	Self-adaptive kNN	BoT-IoT	98%
		ToN-IoT	96.4%
		CICIDS2018	99.9%
[33]	QSVM	UGRansome	99.89%
[21]	Ensemble	CSIC 2012	97.58%
[34]	NERO	Edge-IIoT	99%
Our study	Agentic AI-driven LSTM	ToN-IoT	97.4%
		NSL-KDD	98%

[33], and neural algorithmic reasoning for zero-day detection (NERO) [34]. This competitive performance can be attributed to the LSTM component, which captures temporal attack behavior, and the agentic reasoning mechanism, which combines uncertainty estimation, structural deviation analysis, and adaptive thresholding to improve discrimination between normal, known, and unseen attack traffic.

VI. LIMITATIONS AND FUTURE WORK

Despite the promising results, several limitations remain. Although the proposed framework shows strong performance under cross-dataset distribution shifts, its robustness against adversarial attacks has not yet been evaluated. Future work will examine adversarial transferability, multi-feature perturbation, and adaptive threshold poisoning attacks. In addition, the current framework uses agent-oriented orchestration and adaptive thresholding but does not yet include full autonomous policy learning or continuous environment interaction. Future work will therefore extend the framework with reinforcement learning, incremental learning, and real-time environment-aware adaptation.

VII. CONCLUSION

This study proposed an agentic AI-driven framework for reliable zero-day cyber threat detection to support secure future economies. The framework integrates epistemic uncertainty estimation, embedding-based structural deviation analysis, and adaptive threshold regulation to enable self-regulated and interpretable decision-making beyond conventional closed-set intrusion detection. The experimental results on NSL-KDD and ToN-IoT show that the proposed method improves zero-day detection while maintaining strong performance for known attack classes, low FARs, and stable predictive confidence. The ablation results further confirm that adaptive threshold regulation plays a key role in improving the balance between detection sensitivity and false alarms. Cross-dataset evaluation demonstrates the framework's generalization capability under distribution shift, while the interpretability analysis shows that decisions are supported by meaningful traffic features rather than noisy or arbitrary patterns. Overall, the proposed framework provides a robust, adaptive, and explainable solution for zero-day intrusion detection. Future work will focus on extending the framework with full incremental learning, broader real-time deployment, and validation across more diverse cyber-physical and IoT security domains.

ACKNOWLEDGMENTS

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/35975).

CONFLICT OF INTEREST STATEMENT

The author(s) declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. Al Siam *et al.*, "Securing the unseen: A comprehensive exploration review of AI-powered models for zero-day attack detection," *Expert Syst.*, vol. 43, no. 3, p. e70217, 2026.

[2] S. A. Alansary *et al.*, "Emerging AI threats in cybercrime: A review of zero-day attacks via machine, deep, and federated learning," *Knowl. Inf. Syst.*, vol. 67, no. 11, pp. 10951–10987, 2025.

[3] K. N. Karaca and A. Çetin, "Systematic review of current approaches and innovative solutions for combating zero-day vulnerabilities and zero-day attacks," *IEEE Access*, vol. 13, pp. 102071–102091, Jun. 2025.

[4] P. R. Rajgopal, "Agentic AI for autonomous cybersecurity threat mitigation," in *2025 2nd International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, 2025, pp. 1–7.

[5] T. Vaiyapuri and K. Murugesan, "Integrating terrestrial and non-terrestrial networks in 6G: A review of architectures, AI-driven techniques, and sustainability strategies," *Int. J. Electr. Electron. Eng. Telecommun.*, vol. 15, no. 1, pp. 1–19, 2026.

[6] X. Yang, E. Howley, and M. Schukat, "Agent-based dynamic thresholding for adaptive anomaly detection using reinforcement learning," *Neural Comput. Appl.*, vol. 37, no. 23, pp. 18775, 2025.

[7] T. B. Ogunseyi *et al.*, "Performance analysis of explainable deep learning-based intrusion detection systems for IoT networks: A systematic review," *Sensors*, vol. 26, no. 2, p. 363, 2026.

[8] D. Danang and Z. Mustofa, "CLSTMNet architecture: A CNN–LSTM-based hybrid deep learning model for DDoS attack detection and mitigation in network security," J. Artif. Intell. Technol., vol. 6, pp. 207–214, 2026.

[9] M. G. Vishwanath *et al.*, "Feature-optimized intrusion detection based on a hybrid spiking neural network for the internet of things," *J. Artif. Intell. Technol.*, vol. 6, pp. 52–63, 2026.

[10] K. Alam *et al.*, "Adaptive defense: Zero-day attack detection in nids with deep reinforcement learning," *IEEE Access*, vol. 13, pp. 116345–116361, Jul. 2025.

[11] Y. Medjadba, H. Drid, and M. Rahouti, "Intrusion detection in Software-Defined Networking using hybrid Bayesian model averaging for reliable uncertainty quantification," *Comput. Netw.*, vol. 269, p. 111436, 2025.

[12] R. N. Anaedevha, A. G. Trofimov, and Y. V. Borodachev, "Uncertainty-calibrated hierarchical Gaussian processes for intrusion detection with multi-scale temporal modeling," *Neurocomputing*, vol. 677, p. 133105, May 2026.

[13] D. Krishnan, S. Singh, and V. Sugumaran, "Explainable AI for zero-day attack detection in IoT networks using attention fusion model," *Discov. Internet Things*, vol. 5, no. 1, p. 83, 2025.

[14] N. Kshetri, "Transforming cybersecurity with agentic AI to combat emerging cyber threats," *Telecomm. Policy*, vol. 49, no. 6, p. 102976, Jul. 2025.

[15] A. Jain, R. Bagoria, and P. Arora, "An intelligent zero-day attack detection system using unsupervised machine learning for enhancing cyber security," *Knowl. Based. Syst.*, vol. 324, p. 113833, 2025.

[16] E. K. Boahen and A. S. Shahraki, "ZAD-ML: Dual-layer learning for zero-day attack detection in multivariate time series," *Future Gener. Comput. Syst.*, vol. 180, p. 108422, 2026.

[17] A. Kutlimuratov *et al.*, "A lightweight cascade-based farmework for real-time zero-day attack detection," *Computers*, vol. 15, no. 3, p. 174, 2026.

[18] P. R. Agbedanu *et al.*, "A scalable approach to internet of things and industrial internet of things security: Evaluating adaptive self-adjusting memory k-nearest neighbor for zero-day attack detection," *Sensors*, vol. 25, no. 1, p. 216, 2025.

[19] J. Wilkie *et al.*, "A novel contrastive loss for zero-day network intrusion detection," IEEE Trans. Netw. Serv. Manag., vol. 23, pp. 2064–2076, 2026.

[20] M. Altulyan, K. Murugesan, and T. Vaiyapuri, "Agentic intelligence-driven visual analytics with human-in-the-loop for zero-day attack detection towards a secure future economy," Int. J. Data and Netw. Sci., vol. 10, 2026.

[21] V. Babaey and H. R. Faragardi, "Detecting zero-day web attacks with an ensemble of LSTM, GRU, and stacked autoencoders," *Computers*, vol. 14, no. 6, p. 205, 2025.

[22] A. Gurram, "Generative AI for enhanced cybersecurity: Building a zero-trust architecture with agentic AI," *World J. Adv. Eng. Technol. Sci.*, vol. 15, no. 1, pp. 2380–2396, 2025.

[23] A. Bandi *et al.*, "The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges," *Future Internet*, vol. 17, p. 404, 2025.

[24] M. A. Anvari, D. Rahmati, and S. Kumar, "t-Distributed stochastic neighbor embedding," in Dimensionality Reduction in Machine Learning, S. Chakravert, K. Parand, and J. A. Rad, Ed., Cambridge, MA, USA: Elsevier, 2025, ch. 7, pp. 187–207.

[25] P.-C. Cimpoesu *et al.*, "A t-SNE-based embedding for transfer optimisation with non-overlapping design variables: P.-C. Cimpoesu *et al.*," Struct. Multidiscip. Optim., vol. 68, no. 3, p. 57, 2025.

[26] B. Vijetha, "Agentic intelligence for unified cyber defense: A self-adaptive framework for threat detection across cloud, edge, and IoT systems," *IEEE Access*, vol. 14, pp. 5104–5118, 2026.

[27] M. Alkasassbeh *et al.*, "A self-adaptive intrusion detection system for zero-day attacks using deep Q-networks," *IEEE Access*, vol. 13, pp. 174280–174296, Oct. 2025.

[28] M. Tavallaee *et al.*, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, 2009, pp. 1–6.

[29] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Sustain. Cities Soc.*, vol. 72, p. 102994, Sep. 2021.

[30] H. Rumapea, D. R. Manalu, and Y. Y. P. Rumapea, "Interpretable Deep Learning for Enhanced AI Trust and Clarity," *J. Artif. Intell. Technol.*, vol. 5, pp. 345–353, 2025.

[31] A. Touré *et al.*, "A framework for detecting zero-day exploits in network flows," *Comput. Netw.*, vol. 248, p. 110476, 2024.

[32] H. Hindy *et al.*, "Utilising deep learning techniques for effective zero-day attack detection," *Electronics (Basel)*, vol. 9, no. 10, p. 1684, 2020.

[33] S. J. Nhlapo, E. N. Mutombo, and M. N. W. Nkongolo, "Parameterised quantum SVM with data-driven entanglement for zero-day exploit detection," *Computers*, vol. 14, no. 8, p. 331, 2025.

[34] A. Rizzardi *et al.*, "NERO: Neural algorithmic reasoning for zero-day attack detection in the IoT: A hybrid approach," *Comput. Secur.*, vol. 142, p. 103898, 2024.