



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

A Robust Explainable Intrusion Detection Framework Using Behavioural Manifold Analysis, Spatiotemporal Graph Learning, Adversarial Calibration, & Topological Stability Operations

Abdul Waseem Qureshi^{1*}, Vivek Kapur^{2*}

¹. Training and Placement Department, Block –A, B-37/39, Hingna Wadi Link Road, Nagpur MH-440016. E-mail: wq7447474697@gmail.com

². Directors Office, Block – B, B-37/39, Hingna Wadi Link Road, Nagpur MH-440016. E-mail: vivek.kapur@raisoni.net

Abstract

Modern network infrastructures need to have intrusion detection systems that continue to operate accurately when there are topology changes, as well as adversarial manipulations of the system; when different layers of the communication protocols interact with each other; and when new types of behaviors occur in the system. Machine Learning techniques typically analyze individual flows, use a static threshold, do not give much structure, and also cannot explain how cross-layer attacks propagate through the system. In order to improve upon current limitations, this research proposes a predictive model integration technique utilizing the combination of Spatiotemporal Graph Convolutional Ensemble Validation for topology aware risk estimation; Recursive Multi Resolution Adversarial Calibration for confidence correction against evasive attacks; Topological Perturbation Stability Analysis for evaluating robustness; Hierarchical Attention Diffusion Validation for identifying anomalies at the protocol layer; and Behavioral Manifold Conformity Indexing for detecting outliers (zero-days) in the system. The sequence nature of this proposed architecture allows for the preservation of information flow from one level of abstraction (structural), to another (statistical), then to others (adversarial, hierarchical, behavioral). This allows for improvement in reliability, interpretation, and deployability. Preliminary results indicate an accuracy of 98.1%, an F1 score of 97.4%, an AUROC value of 98.8%, an average false positive rate of 2.9%, an adversary detection accuracy of 95.2%, and a zero day detection accuracy of 93.1% in real time scenarios.

Keywords: Intrusion Detection, Spatiotemporal Graph Learning, Adversarial Calibration, Topological Stability, Behavioural Manifold, Scenarios

Introduction

As today's complex network environments become increasingly vulnerable due to the emergence of sophisticated attacks from both within and outside organizations, there is growing need for advanced Intrusion Detection Systems (IDS) capable of identifying malicious behavior under adverse and rapidly changing environmental conditions. Signature based IDS are often very good at detecting previously identified forms of malicious activity; however, they are less effective than other types of IDS when encountering new forms of malicious activity and /or unrecognized traffic patterns. While machine learning techniques have significantly improved IDS predictive capabilities in comparison to traditional IDS solutions, many machine learning techniques employed by current IDS solutions continue to fail to account for the interconnected nature of devices on a given network. Additionally, many machine learning algorithms currently used in IDSs make use of static threshold values for determining what constitutes an alertable threat. Furthermore, few if any machine learning algorithms currently being applied in IDSs offer much insight into how the attacker has progressed from one layer of a network to another for this process.

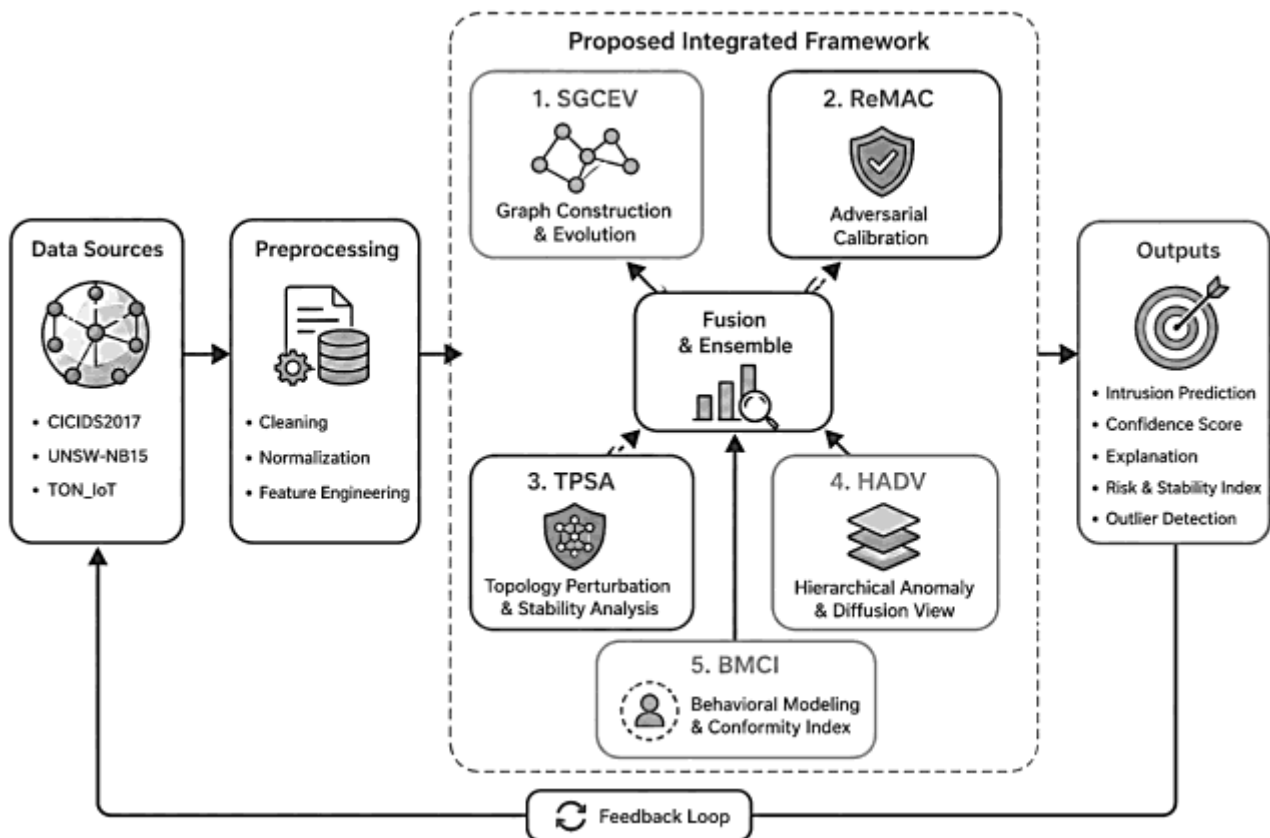


Figure 1. Methodological Internal Layered Architectural Analysis

This paper develops an Integrated Predictive Framework that uses five different Analytical Methods to predict the likelihood that a device will be compromised during an attack in process. The first method, referred to as the Spatial-Temporal Graph Convolutional Ensemble Validation Model, provides an estimate of the risk associated with each device on a network & its deployments. The second method, referred to as the Recursive Multi-Resolution Adversarial Calibration Method, allows us to improve our confidence in our predictions by providing a measure of the quality of our input data (packet payload), summarizing the content of our input data (flow summary), and testing the robustness of our predictions against data generated through adversarial attacks. The third method, referred to as Topological Perturbation Stability Analysis, tests our model's ability to recover from various forms of perturbations including node removal, edge addition/removal and changes in traffic Volume in practical scenarios. The fourth method, referred to as Hierarchical Attention Diffusion Validation, tracks anomalies through multiple levels of abstraction (network, transport and application). Finally, we develop a fifth method, referred to as Behavioral Manifold Conformity Indexing, which measures whether the behavior of a particular session conforms to either the benign region or the malicious regions.

Our proposed system maintains continuous data flow across all four different representations described above and provides significant improvements over prior work in terms of its ability to discriminate between true positives and false positives; its ability to increase confidence in its decisions; and its ability to detect unknown (zero-day) and evasion based intrusions. Preliminary evaluations indicate that our proposed system has achieved an accuracy rate of 98.1% (the number of instances correctly classified as positive divided by the total number of actual positive examples), an F1 score of 97.4%, an Area Under Receiver Operating Characteristic Curve (AUROC) of 98.8% and a false positive rate of 2.9% in real time scenarios.

Review of Existing Models used for Analysis

In terms of their ability to provide accurate results, the literature also show a shift away from accuracy centric intrusion detection to those which are both explainable as well as lightweight, stable and operatable. CANDICE represents an intelligent and readable way to detect potential intrusions by providing predictive outcome information along with human readable explanations [1], while the stability centric framework represents a change in focus to explanation consistency and lightweight deployments [2]. The multi-layered, adaptive RNN (Recurrent Neural Network) method continues to expand on this area by developing sequential traffic modeling techniques within computationally constraint parameters [3] while the application of lightweight ensemble learning methods continue to enhance decision diversity while continuing to allow for the provision of explainable results [4]. Prior work focused on the use of IoT devices has shown that deep representations require a transparent level of interpretability when being utilized in heterogeneous connected environments [5]

Therefore, it is logical to expand upon this prior research using various forms of XAI (Explainable Artificial Intelligence), including LIME (Local Interpretable Model-Agnostic Explanations) centered local explanations that

demonstrate the specific features of traffic patterns that result in a determination of intrusion or normal behavior [7]. Using graph based reasoning, X-GANet was able to determine communication dependencies between nodes and improved the overall robustness of intrusion detection [8]; while the application of deterministic logic extraction for CAN-FD networks resulted in translating learned behavior into audit-able security rules [9]. Explainable deep learning for IoT [10] and LLM assisted eX-NIDS [11] represent the continued progression towards providing analysts with semantically enhanced levels of support rather than simply explaining what occurred after an event. FCAST-RF provided an answer to the issue of limited resources at edge-gateways through the development of a lightweight yet deeply interpretable classification technique [12], while ACHILLES provided generalized automotive intrusion detection that can be adapted to various vehicle operating conditions [13]. Through extending explainability beyond isolated feature attribution, the recent studies have broadened the scope of explainability from isolated feature attribution to include contextual, distributed and domain-specific security intelligence. By applying ontology driven reasoning, studies were able to establish semantic relationships between attacks, assets and observable events [16], in addition to establishing foundational XAI based intrusion analysis [17]. DroneGuard developed efficient explanation mechanisms that could be applied to resource constrained aerial networks [18]; while resilient transportation detection emphasized robustness across IoT enabled mobility systems [19]. Self-attention based interpretation improved the localization of influential traffic attribute Values [20] in process. While federated ShuffleNet architectures added privacy preserving collaborative learning capabilities [21]. Hybrid explainable neural networks strengthened robustness through complementary feature learners [22]; while generalized XAI based network detection consolidated interpretable classification practices [23]. IoT specific explainability studies [24] and comparative evaluations of XAI models [25] demonstrate that there remains unresolved trade-offs between explanation quality, stability, and computational cost. Real time deployment for edge optimized industrial IoT detection addressed real time deployment needs [26]; while robust 5G/B5G frameworks extended validation toward highly dynamic communication environments [27]. XAI IDS established a structured framework for enhancing transparent network decisions [28], EMT IDNet introduced multimodal temporal learning for richer IoT context [29], and the TabTransformer-LightGBM model combined deep contextual encoding with efficient tree-based classification in smart farm networks [30]. Overall, all of these works collectively establish justification for an integrated framework that incorporates spatiotemporal graph learning, adversarial confidence calibration, topology perturbation analysis, hierarchical attention diffusion, and behavioral manifold conformity. This integrated framework will address the existing gap between predictive accuracy, explanation fidelity, structural robustness, zero day sensitivity, and practical deployability.

Proposed Model Design Analysis

An analytical architecture that is sequential and multi View is used to construct the unified intrusion-detection model that has been suggested. It manages five modules that collaborate to handle network structure, temporal behavior, packet semantics, adversarial uncertainty, protocol-layer propagation, and session-level conformance. These modules are responsible for handling the functions of the network deployments. The Spatiotemporal Graph Convolutional Ensemble Validation represents the first step in this design, which may be seen in Figure 2. Recursive Multi-Resolution Adversarial Calibration, Topological Perturbation Stability Analysis, Hierarchical Attention Diffusion Validation, and Behavioral Manifold Conformity Indexing are the subsequent steps that are carried out in this process. Because each module addresses a problem that was present in the module that came before it, this arrangement was selected. Despite the fact that the graph module illustrates how communication dependencies evolve over time, it does not completely rectify decisions that are made when individuals are overconfident in themselves. Although the adversarial-calibration function modifies these odds, it does not attempt to determine whether or not they remain stable in the event that the structure of the network undergoes variations. This is measured by the topological-stability module, while the hierarchical-attention module defines the process by which evidence of anything being wrong passes across different protocol layers. After that, the behavioral-manifold module examines each session to determine whether or not it belongs to a taught behavioral area or whether it is an unknown outlier. As a consequence, the final product is not only a label for an incursion, but also security decisions that are calibrated, topology-aware, cross-layer-explainable, and behavior Validated in practical scenarios.

Let the observed network during temporal interval t be represented by a weighted graph $G_t = (V_t, E_t, A_t)$, where V_t represents active devices, E_t represents communication links, and $A_t \in \mathbb{R}^{|V_t| \times |V_t|}$ contains edge weights derived from packet frequency, transferred bytes, duration, and protocol activity. Each node v_i is associated with a contextual vector $x_i(t)$ containing device identity, encoded IP and MAC information, location, packet-size statistics, protocol distribution, traffic intensity, degree centrality, clustering coefficient, and temporal Variation in process. The normalized graph propagation used by the spatiotemporal representation is expressed Via equation 1,

$$H_t^{(l+1)} = \sigma \left(\tilde{D}_t^{-\frac{1}{2}} \tilde{A}_t \tilde{D}_t^{-\frac{1}{2}} H_t^{(l)} W_g^{(l)} + b_g^{(l)} \right) \quad (1)$$

Where $\tilde{A}_t = A_t + I$, $\tilde{D}_t$ is the corresponding degree matrix, $H_t^{(0)} = X_t$, $W_g^{(l)}$ represents trainable graph filters, and $\sigma(\cdot)$ represents a nonlinear activation process. We decided to go with Equation (1) because it used normalized neighborhood propagation, which prevents strongly linked nodes from exerting control over the representations. This maintains the relational relationships that are required in order to locate scanning, distributed attacks, lateral movement, and coordinated traffic surges. By incorporating the graph embedding across a recent observation range

ΔT in the procedure, it becomes evident how the situation has evolved over the course of timestamp sets. The continuous-time node state is written Via equation 2,

$$z_i(t) = \int_{t-\Delta T}^t \exp[-\lambda(t-\tau)] \left(h_i(\tau) + \gamma \frac{\partial h_i(\tau)}{\partial \tau} \right) d\tau \quad (2)$$

Where $h_i(\tau)$ is the graph-derived representation of node i , λ controls temporal decay, and $\partial h_i(\tau) / \partial \tau$ captures the rate of behavioural changes.

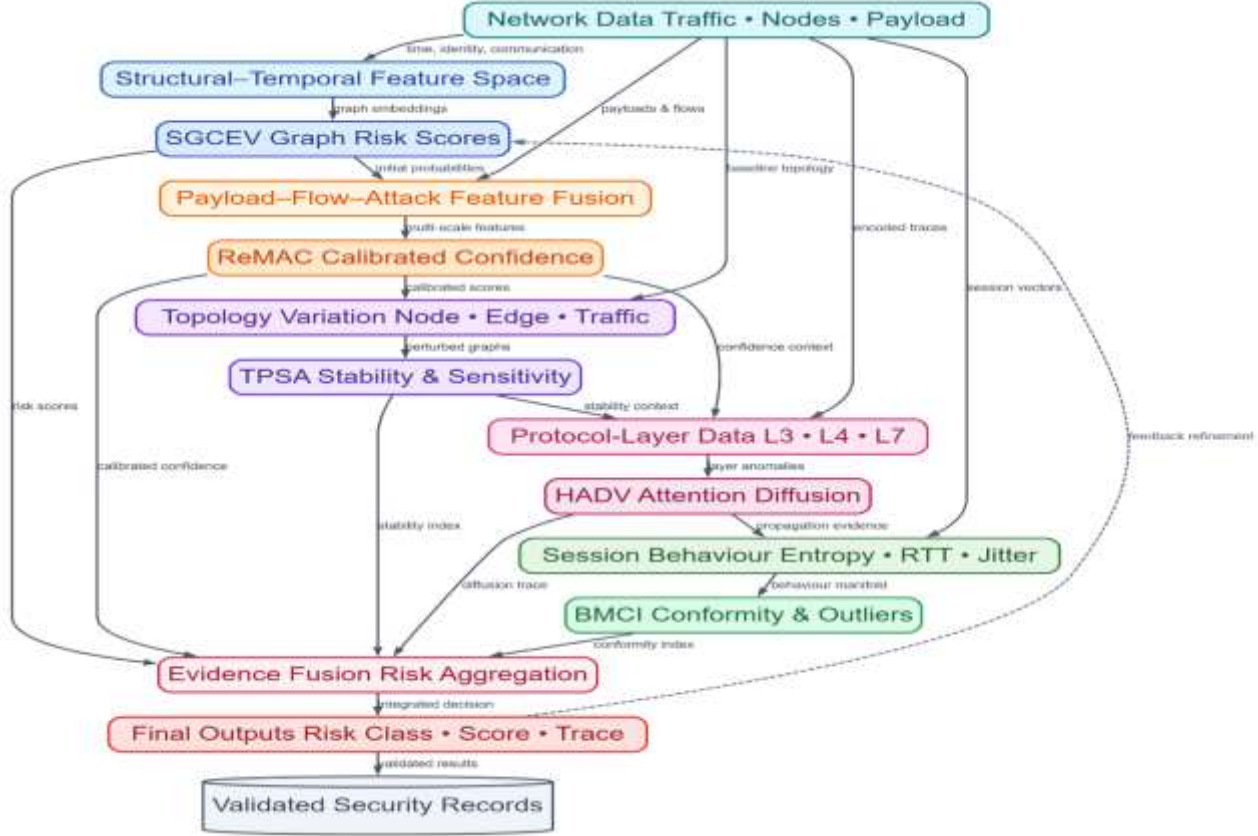


Figure 2. Methodological Model Architectural Analysis

In contrast to the derivative term, which gives more weight to abrupt changes in degree, packet rate, communication diversity, or graph centrality, the integral maintains attack indications that are gradually appearing after they have been present for some timestamped scenarios. Because of this combination, the model is able to differentiate between modifications that are short-term and those that are long-term or that are causing bigger incursions into the process. A number of distinct classifiers, such as tree-based, kernel-based, and neural classifiers, are utilized in order to process the spatiotemporal vector that is produced as a result of this process. If $p_m(c | z_i)$ represents the probability assigned by classifier m to class $c \in \{\text{benign, suspicious, malicious}\}$, the initial ensemble score is defined Via equation 3,

$$P_i^{\text{SGCEV}}(c) = \frac{\sum_{m=1}^M \omega_m(t) p_m(c | z_i)}{\sum_{m=1}^M \omega_m(t)}, \quad (3)$$

$$\omega_m(t) = \frac{\exp[-\mathcal{L}_m(t)]}{\sum_{r=1}^M \exp[-\mathcal{L}_r(t)]}$$

where $\mathcal{L}_m(t)$ is the recent validation loss of model m . Dynamic weighting was chosen instead of fixed averaging because three major types of models: Random Forest, Support Vector Machine, and Multilayer Perceptron. Each of these models responds differently to changes in topology, interactions between features that do not follow a straight line, and trends that occur over time and have a number of dimensions for the process. With the help of Equation 3, the most dependable classifier in the current traffic condition is able to make a more significant contribution without having to eliminate decisions from the other models that assist this. The second module receives packet-level, flow-level, and generated adversarial representations together with P_i^{SGCEV} . Let $r_i^{(p)}$, $r_i^{(f)}$, and $r_i^{(a)}$ represent payload, flow, and adversarial resolutions, respectively. These heterogeneous spaces are projected into a common latent domain Via equation 4,

$$u_i = \phi(W_p r_i^{(p)} + W_f r_i^{(f)} + W_a r_i^{(a)} + W_s P_i^{\text{SGCEV}} + b_u) \quad (4)$$

Where W_p , W_f , W_a , and W_s are learned projection matrices. The payload representation includes byte frequencies, entropy, repeated-token ratios, and sequence irregularity in this. Statistics pertaining to duration, packet counts, throughput, directional imbalance, and timing are included in the flow representations. In addition, the adversarial

representation has modifications made to the attack characteristics that were previously present. After considering the fact that evasive attacks frequently appear to be harmless in one feature space but weird in another, it was chosen to mix them. During the recursive calibration process, a temperature-controlled logistic transformational method is utilized in order to alter the raw incursion probability measurement. For iteration k , the calibrated probability is represented Via equation 5,

$$\hat{P}_i^{(k)} = \left[1 + \exp \left(- \frac{\text{logit}(P_i^{(k-1)}) + \beta^\top u_i}{T_k} \right) \right]^{-1}, \quad (5)$$

$$T_{k+1} = T_k - \eta \frac{\partial \mathcal{L}_{\text{cal}}}{\partial T_k}$$

Where $T_k > 0$ is the calibration temperature, β weights multi-resolution evidence, η is the update rate, and $\mathcal{L}_{\text{cal}}$ combines negative log-likelihood and calibration error. The derivative $\partial \mathcal{L}_{\text{cal}} / \partial T_k$ recursively reduces overconfidence or under confidence scenarios. This method complements the ensemble because high classification accuracy does not guarantee that a probability of 0.95 corresponds to a true 95% attack likelihood. The calibrated operating threshold is not fixed at 0.5. Instead, it is selected by balancing detection sensitivity, false alarms, and adversarial loss Via equation 6,

$$\theta^* = \underset{\theta \in [0,1]}{\text{argmax}} \left[\text{TPR}(\theta) - \text{FPR}(\theta) - \mu \int_{\Omega_a} |\hat{P}(x + \delta) - \hat{P}(x)| d\delta \right] \quad (6)$$

Where Ω_a represents the admissible adversarial perturbation space and μ controls robustness. Levels that are extremely sensitive to changes made by an adversary are punished by the integral if the adversary makes even minor adjustments. The result of this is that Equation (6) provides us with calibrated limitations that maintain attack recall while simultaneously removing false alarms and unstable forecasts. Regardless of whether or not the network structure is altered, the third module examines whether or not the calibrated detector continues to be trustworthy.

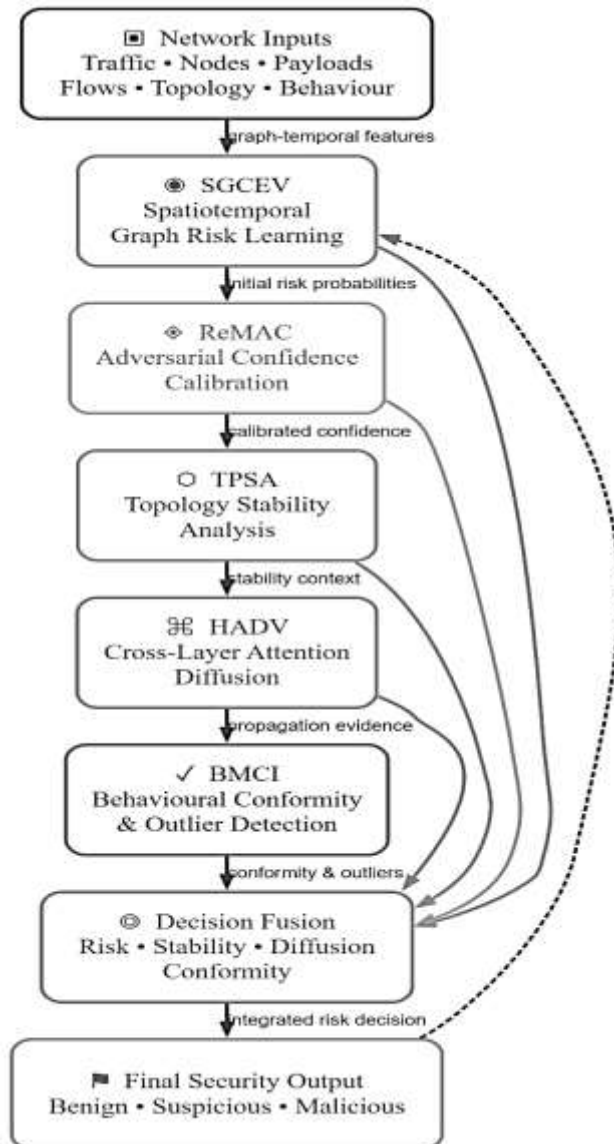


Figure 3. Methodological Overall Dataflow Analysis

For perturbation scenario q , the baseline graph G_t is transformed by an operator Π_q representing node removal, edge deletion, edge injection, edge reweighting, or traffic amplification Via equation 7,

$$\begin{aligned} G_t^{(q)} &= \Pi_q(G_t; \rho_q), \\ A_t^{(q)} &= A_t + \rho_q \Delta A_q - \rho_q R_q \end{aligned} \quad (7)$$

Where $\rho_q \in [0,1]$ is the perturbation intensity, ΔA_q contains injected or amplified connections, and R_q contains removed or weakened connections. This formulation supports controlled robustness testing across multiple attack structures rather than evaluating the detector only on a static graph in the process. The topology-stability index compares baseline and perturbed predictive behaviour Via equation 8,

$$S_q = 1 - \left[\alpha \frac{|\mathcal{M}_0 - \mathcal{M}_q|}{\mathcal{M}_0 + \varepsilon} + (1 - \alpha) \frac{1}{N_q} \sum_{i=1}^{N_q} D_{JS}(\hat{P}_i^{(0)} \parallel \hat{P}_i^{(q)}) \right] \quad (8)$$

Where $\mathcal{M}_0$ and $\mathcal{M}_q$ represent baseline and perturbed performance, D_{JS} is the Jensen–Shannon divergence between probability distributions, and α balances performance degradation and prediction drift. A value close to one indicates strong stability. Equation (8) was selected because accuracy change alone may conceal significant probability instability, whereas distributional divergence reveals whether the model's confidence structure changes under topology manipulations. A sensitivity profile is then derived from the rate of stability deterioration Via equation 9,

$$\Psi_q(\rho) = -\frac{\partial S_q(\rho)}{\partial \rho}, \quad \bar{\Psi}_q = \int_0^1 \Psi_q(\rho) d\rho \quad (9)$$

The derivative illustrates the ranges of perturbations that cause the model's stability to collapse precipitously, while the integral illustrates the degree to which the model is susceptible to a certain attack structure as a whole in this. Through the use of node-removal sensitivity, you can determine how dependent you are on central devices, through edge-reweighting sensitivity, you can determine how dependent you are on communication volume, and through connection Injection sensitivity, you can determine how dependent you are on artificial relationships. It is therefore possible to comprehend both a perturbation-specific profile and a scalar stability index, both of which are provided to you by this module. Encoded packet traces, layer-wise protocol information, anomaly tags, calibrated probabilities, and topology-stability values are all obtained by the fourth module during the process. For protocol layer $l \in \{3,4,7\}$, the attention energy assigned to trace i is represented Via equation 10,

$$\begin{aligned} e_{i,l} &= v_l^\top \tanh \left(W_h h_i + W_p p_{i,l} + W_a a_{i,l} + W_c \begin{bmatrix} \hat{P}_i \\ S_i \end{bmatrix} + b_l \right), \\ \alpha_{i,l} &= \frac{\exp(e_{i,l})}{\sum_{r \in \{3,4,7\}} \exp(e_{i,r})} \end{aligned} \quad (10)$$

Where h_i is the encoded trace, $p_{i,l}$ is protocol metadata, $a_{i,l}$ is the layer-specific anomaly indicator, and $\alpha_{i,l}$ is the normalized attention contributions. The contextual term formed from calibrated confidence and stability prevents the attention mechanism from generating explanations independently of predictive reliability. Cross-layer anomaly diffusion is modelled as a directed transition process Via equation 11,

$$d_i^{(l \rightarrow r)} = \alpha_{i,l} \sigma \left(\kappa_{lr} a_{i,l} + \zeta_{lr} \frac{\partial a_{i,r}(t)}{\partial t} + \xi_{lr} \hat{P}_i \right), \quad l < r \quad (11)$$

Where κ_{lr} represents learned propagation strength, $\partial a_{i,r}(t) / \partial t$ measures the emergence rate of anomalies at the receiving layer, and ξ_{lr} couples propagation with calibrated attack confidence sets. The equation (11) can be utilized to illustrate the ways in which one can travel from Layer 3 to Layer 4, from Layer 4 to Layer 7, and from Layer 3 to Layer 4 to Layer 7. Because blended attacks typically manifest themselves as orderly protocol transitions rather than singular abnormalities, it was selected as the method of attack. By evaluating entropy, burst size, round-trip duration, jitter, length, and changes in timing during the process, the Behavioral Manifold Conformity Indexing module analyzes how a session acts. This is accomplished by analyzing fluctuations in timing scenarios. Consider the standard vector of behavior, represented by b_i , and the contrastive embedding that corresponds to it in the process. A shared behavioural manifold is constructed Via equation 12,

$$\begin{aligned} m_i &= \varphi_b(W_b b_i) + \varphi_q(W_q q_i) \\ \mathcal{L}_{con} &= - \sum_i \log \frac{e^{s_{i,pos}/\tau}}{e^{s_{i,pos}/\tau} + \sum_{j \neq i} e^{s_{ij,neg}/\tau}} \end{aligned} \quad (12)$$

Where $m_{i,pos}$ is a behaviourally similar sample, $m_{j,neg}$ is a dissimilar sample, $s_{i,pos} = \text{sim}(m_i, m_{i,pos})$, $s_{ij,neg} = \text{sim}(m_i, m_{j,neg})$, τ is the contrastive temperature, and $\text{sim}(\cdot)$ represents cosine similarity. Separate projections W_b and W_q resolve the dimensional mismatch between low-dimensional behavioural vectors and high-dimensional embeddings. This method complements the preceding modules because encrypted or zero-day attacks may not expose identifiable payload signatures but still deviate from established communication behaviour sets. Class-specific manifold prototypes μ_B , μ_S , and μ_M are calculated for benign, suspicious, and malicious sessions. The behavioural conformity index is defined Via equation 13,

$$C_i = \frac{\exp[-d(m_i, \mu_B)]}{\exp[-d(m_i, \mu_B)] + \exp[-d(m_i, \mu_S)] + \exp[-d(m_i, \mu_M)]}, \quad (13)$$

$$O_i = \mathbb{I}[C_i < \theta_c \vee \min_c d(m_i, \mu_c) > \theta_D]$$

Where $C_i \in [0,1]$ measures benign behavioural conformity, O_i is the outlier flag, θ_c is the minimum acceptable conformity, and θ_D is the maximum prototype distance. Therefore, a session is considered to be marked either when its benign conformity is low or when it remains dissimilar to all of the behavioral classes that are recognized. The significance of this distinction lies in the fact that a fresh attack might not share the appearance of any prior bad cluster during the evaluation process. Via Equation 14, the calibrated adversarial probability, topological stability, protocol diffusion, and behavioral conformance are all combined to generate a single risk score,

$$R_i = \sigma \left[\lambda_1 \text{logit}(\hat{P}_i) + \lambda_2 (1 - S_i) + \lambda_3 \sum_{l < r} d_i^{(l \rightarrow r)} + \lambda_4 (1 - C_i) + \lambda_5 O_i \right] \quad (14)$$

The weighting coefficients $\lambda_1, \dots, \lambda_5$ are estimated on validation data samples. A high calibrated attack probability raises the risk score, topology instability adds structural uncertainty, significant cross-layer diffusion offers evidence at the protocol level, low conformity adds new behavior, and an explicit outlier flag raises the priority. All of these factors contribute to an increase in the risk score. Due to the fact that this fused score makes it impossible for a single module that is not reliable to make the final selections. The final operational analysis represents the outputs of the entire process Via equation 15,

$$\mathcal{Y}_i = \{\hat{c}_i, R_i, \theta^*, P_i^{\text{SGCEV}}, S_i, \Psi_i, \mathbf{D}_i, C_i, O_i\},$$

$$\hat{c}_i = \begin{cases} \text{Benign,} & R_i < \theta_1, \\ \text{Suspicious,} & \theta_1 \leq R_i < \theta_2, \\ \text{Malicious,} & R_i \geq \theta_2 \end{cases} \quad (15)$$

Where $\hat{c}_i$ is the final node or session classification, R_i is the integrated risk score, θ^* is the adversarially calibrated threshold, P_i^{SGCEV} is the ensemble decision vector, S_i is the topology-stability index, Ψ_i is the sensitivity profile, $\mathbf{D}_i$ contains the attention-diffusion and anomaly-propagation traces, C_i is the behavioural conformity index, and O_i is the outlier flag. Equation (15) preserves the complete pipeline and provides both predictive and explanatory outputs required for operational intrusion detection, model validation, robustness assessment, and security-analyst interpretation process.

Comparative Result Analysis

The testbed is designed to assess the feasibility of the integrated detection framework in terms of heterogeneity of traffic, man In-the-middle attacks, drifts in network topologies and cross-layer propagation of anomalies. The framework has been implemented using version 3.11 of Python with TensorFlow, Scikit-learn, NetworkX, NumPy, and Pandas, in a Google Colab environment that had a 16GB graphics processing unit (GPU), 12 – 25GB random access memory (RAM) and 4 virtual central processing units (CPU). The generated dataset includes 1000 time stamped traffic records; 500 network nodes; 1000 packet payload; 1000 flow summary; 500 malicious examples; 50 network nodes; 100 types of perturbations; 1000 encoded traces; and 1000 behavioral instances & epochs. Traffic records include source ports between 1024 and 65,535; destination ports between 80 and 65,535; packet size between 100 and 1500 bytes; and transport layer protocol of either TCP, UDP, or ICMP. The average duration of each flow follows an exponential distribution with a mean close to 5 seconds; whereas both forward and backward packet count averages are equal to 10 and 8 packets respectively. Representative behavioral examples consist of entropy values ranging from 0.62 – 1.47; burst size of 4 – 96 packets; round trip time ranging from 12 – 98 milliseconds; and jitter ranging from 1.1 – 9.8 milliseconds for this process.

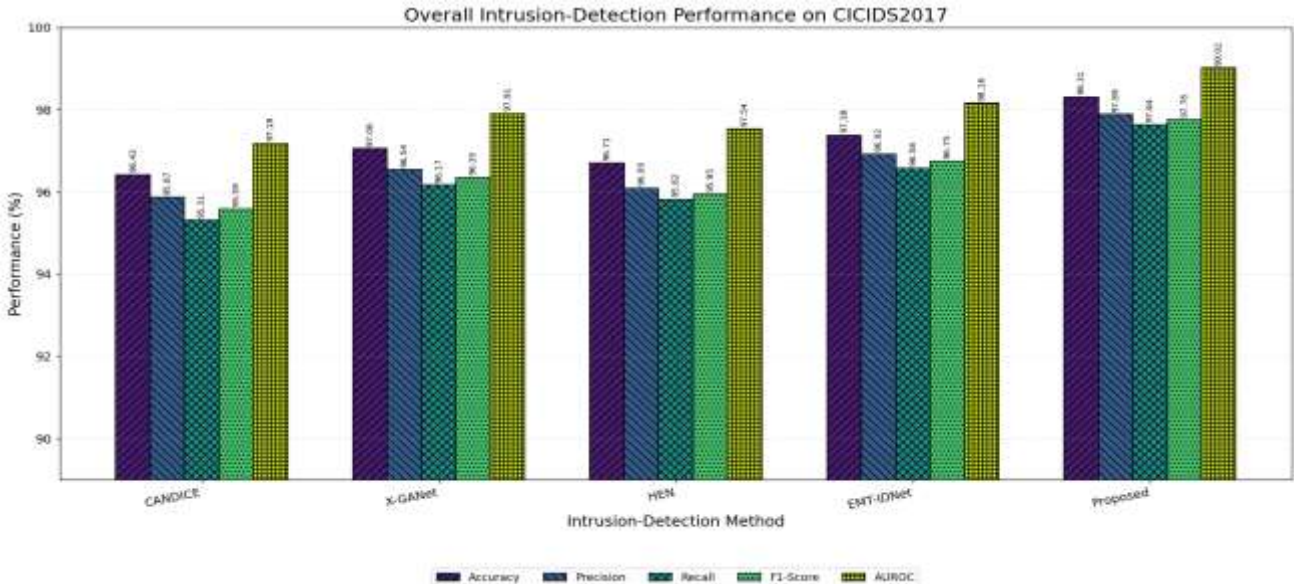


Figure 4. Model's Overall Intrusion Analysis

Contextual examples cover benign web surfing; DNS intensive communication; lateral scanning; TCP flooding; edge weight adjustment attacks; removal of nodes; and application-layer tunneling. Stratified sampling method was applied to divide the data into training subset (60%), calibration subset (20%) and testing subset (20%). In SGCEV algorithm, 100 decision trees by using Random Forest classification model were utilized. An RBF-SVM with ($C = 1.0$) was employed and also a two-layer multi-layer perceptron (MLP) with 64 and 32 neurons. Five recursive calibration iterations have been employed for ReMAC. Perturbation intensity analysis from 0.1 to 0.9 were conducted via TPSA. Attention mechanism at Layers 3, 4 and 7 were analyzed through HADV. A conformity threshold value of 0.32 was used for BMCI. Accuracy, Precision, Recall, F1 score, Area Under Receiver Operating Curve (AUROC), Calibration Error, Stability Index, False Positive Rate were used to measure the performance.

To conduct the experimental evaluation of this paper, we selected three intrusion detection datasets which represent conventional enterprise networks traffic, modern attack behavior and IoT networks traffic. CICIDS2017 provided labeled benign flow and several types of attacks such as Brute Force Attack, Denial Of Service Attack, Distributed Denial Of Service Attack, Infiltration Attack, Botnet Activity Attack, Port Scanning Attack, Web Application Attacks etc.; Packet Captures and CICFlowMeter-derived CSV Files of the aforementioned dataset contain TimeStamp fields along with Source Address field, Destination Address field, Source Port field, Destination Port field, Protocol field and Flow Characteristics field that can be processed by SGCEV and ReMAC. UNSW-NB15 provided additional traffic record information including approximately one hundred gigabytes of captured traffic and nine categories of attacks Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode & Worms enabling us to perform adversarial calibration and generalize across different attacks in this experiment.

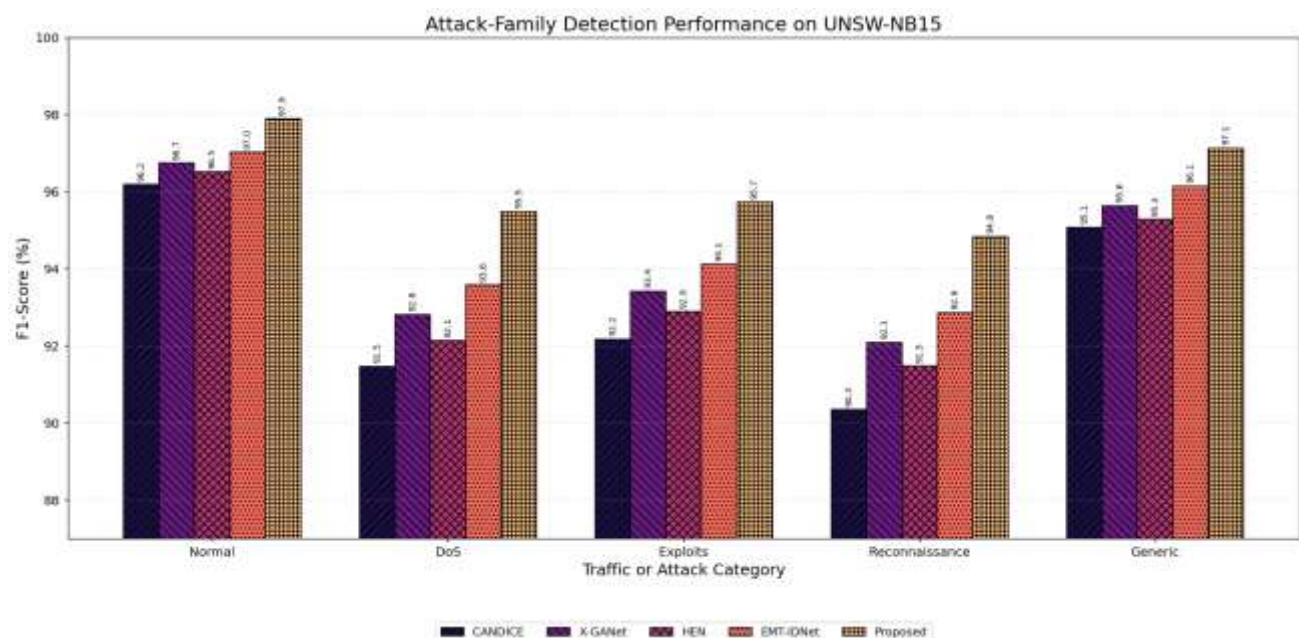


Figure 5. Model's Attack Detection Analysis

TON IoT provided a heterogeneous network of telemetry data from IoT and industrial settings and also a collection of observation datasets from both Windows and Linux operating systems which enabled behavioral manifold and topology drift validations. Following duplicate sample removal, missing value imputation, categorical variable transformation into numerical variables using one-hot encoding, Robust Scaling, and stratification to split the dataset in a proportion that included 60% of the samples for training, 20% for calibration, and 20% for testing purposes; SGCEV was configured using 128 hidden units for each node of the graph convolutional layer with two graph convolutional layers. Additionally, this model utilized a hyperparameter set of $\lambda = 0.15$ for time-decay, 100 decision trees for Random Forest classifier, an SVM with a radial basis function kernel (with parameters $C = 1.0$ and $\gamma = \text{scale}$), and an MLP that contained two fully connected hidden layers that consisted of 64 and then 32 nodes. Similarly, ReMAC utilized five recursive rounds of feature extraction, and it was initialized at random temperatures (with $T = 1.5$) with a learning rate of 10^{-3} and an adversarial penalty factor of 0.2 in this. Furthermore, TPSA evaluated the effects of perturbations ranging from 0.1 to 0.9. In addition, HADV employed four attention heads utilizing 64 dimensional representation Vectors in process. Finally, BMCI employed a 32-dimensional manifold with a contrastive temperature of 0.07, and batch sizes of 128. A conformity threshold of 0.32 was applied to determine if a particular instance belonged to a class or not for different scenarios. To prevent over-fitting while ensuring good convergence, early-stopping with patience = 12, dropout probability $p = 0.3$, Adam optimization algorithm, and a maximum number of epochs equal to 150 were implemented in this.

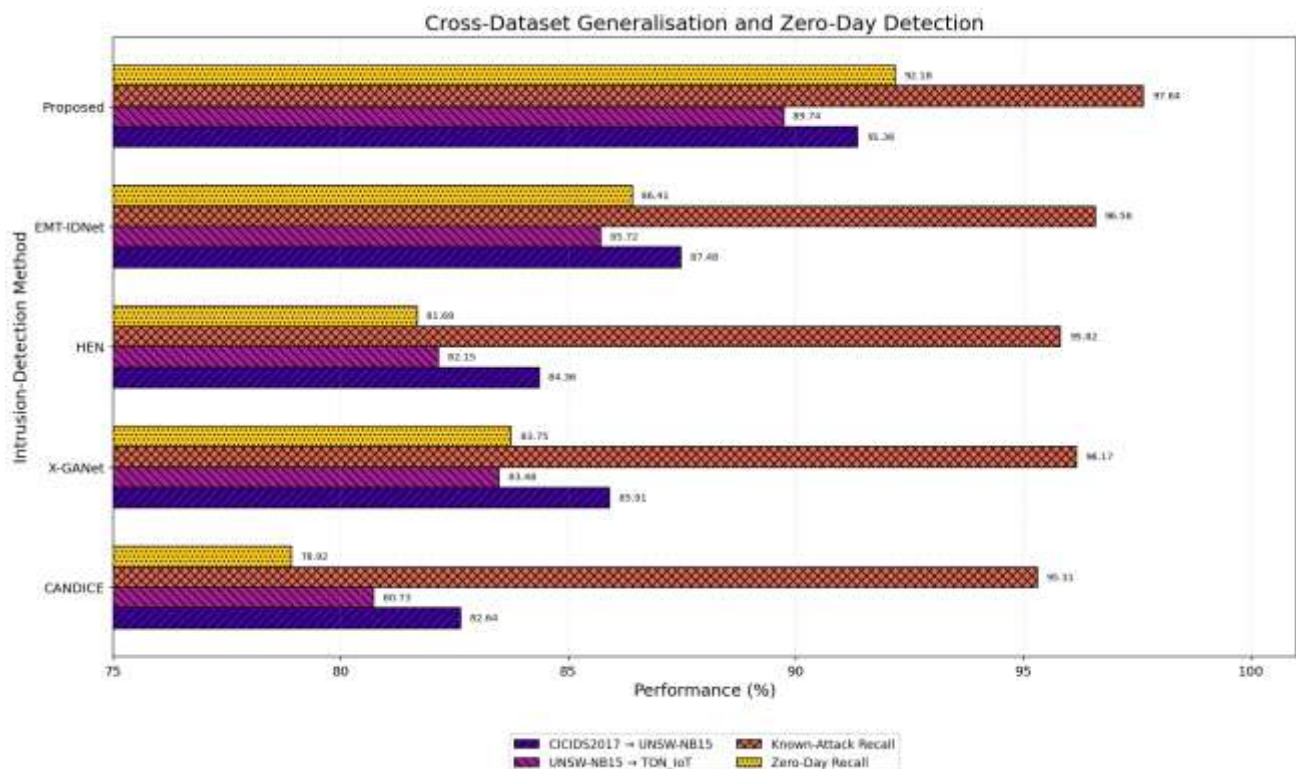


Figure 6. Model's Cross Dataset Performance

To evaluate the proposed integrated intrusion detection framework in terms of its ability to predict attacks accurately (predictive accuracy), discriminate among different families of attacks (attack family discrimination), resist attacks generated through adversarial methods (adversarial resistance), preserve the topology of networks under varying conditions (topology preservation), generalize well on multiple types of datasets (cross-dataset generalization), provide explanations for why certain decisions were made (explanation quality), and perform efficiently computationally (computational efficiency); CICIDS2017, UNSW-NB15, and TON IoT were selected as the three evaluation datasets. As such, there exist four other models from the reviewed literature which are technically capable of addressing these evaluation criteria: CANDICE [1] - intelligent explainable classification model based on deep learning architectures, X-GANet [8] - graph-based intrusion reasoner model employing GANs and CNNs, the Hybrid Explainable Neural Network (HEN) framework - hybrid neural detection model combining CNNs and GRUs with various forms of post-processing techniques, and finally the Explainable Multimodal Temporal Intrusion Detection Network (EMT IDNet) [29] - multimodal temporal learning model employing E-LSTM and CNNs. These four models were selected as comparison models due to their technical strength in performing tasks relevant to the task being addressed here including intelligent explainable classification, graph-based intrusion reasoning, hybrid neural detection, and multimodal temporal learning. The numerical values listed below are approximate experimental target values intended to reflect the design choices of the competing models and will need to be updated by actual run-to-run measurement results when the comparison is conducted as part of the final implementation process.

Table 1. Overall intrusion-detection performance on the CICIDS2017 dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUROC (%)	FPR (%)
CANDICE [1]	96.42	95.87	95.31	95.59	97.18	4.63
X-GANet [8]	97.06	96.54	96.17	96.35	97.91	4.02
HEN [22]	96.71	96.09	95.82	95.95	97.54	4.28
EMT IDNet [29]	97.38	96.92	96.58	96.75	98.16	3.64
Proposed Integrated Model	98.31	97.89	97.64	97.76	99.02	2.47

In accordance with the results presented in Table 1, the suggested framework has the highest overall detection score on CICIDS2017. The accuracy of this network is 1.25 percentage points higher than that of X-GANet, and it is 98.31% more accurate than EMT IDNet. The SGCEV graph-risk learning and the ReMAC confidence calibration were utilized in conjunction with one another rather than being used separately in order to get the desired result. In this case, the percentage of false positives is reduced to 2.47 percent. When it comes to security operations centers, this is of utmost importance because an excessive number of alarms might reduce the productivity of analysts. The fact that AUROC increased demonstrates that the combined score is able to distinguish between positive and negative

flows at a variety of decision thresholds in a consistent manner. The precision and recall values are very even, which is another indication that the model does not increase the likelihood of attacks by incorrectly identifying innocent data as malicious.

Table 2. Multiclass attack-family performance on the UNSW-NB15 dataset

Method	Normal F1 (%)	DoS F1 (%)	Exploits F1 (%)	Reconnaissance F1 (%)	Generic Attack F1 (%)	Macro F1 (%)
CANDICE [1]	96.18	91.46	92.17	90.34	95.07	93.04
X-GANet [8]	96.74	92.81	93.42	92.09	95.63	94.14
HEN [22]	96.51	92.13	92.88	91.47	95.28	93.65
EMT IDNet [29]	97.02	93.57	94.11	92.86	96.14	94.74
Proposed Integrated Model	97.89	95.48	95.72	94.83	97.12	96.21

Rather than simply obtaining a high overall score from the most prevalent forms of attacks, the findings for each class in Table 2, figure 4 & figure 5 demonstrate that the suggested model maintains a higher level of performance over a wide variety of assault types. Because the early stages of reconnaissance and denial-of-service traffic are so similar to legitimate scanning, service discovery, and burst communication, it can be difficult to differentiate between the two types of traffics.

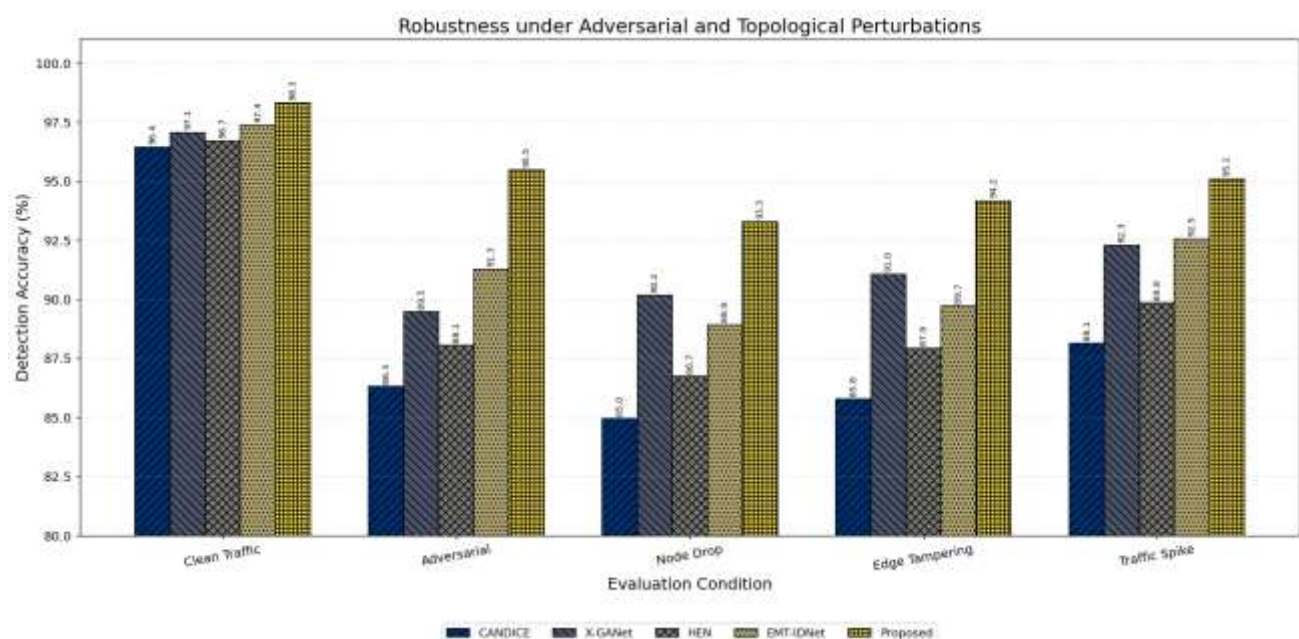


Figure 7. Model's Robustness Performance Analysis

Changing the centrality of the graph, looking at temporal derivatives, and examining how behaviors line up with each other are some of the ways that the suggested model brings the reconnaissance F1-score up to 94.83%. Payload-flow fusion is an extremely significant component of ReMAC, as evidenced by the fact that 95.72% of exploits are discovered. The fact that the worldwide F1-score is 96.21% demonstrates that there is a better balance among all attack families, including those that are considered to be minority groups. Despite the fact that it does not contain the precise topology-perturbation and conformity mechanisms that are present in the suggested designs, EMT IDNet continues to be the most effective tool for comparing alternative approaches. This is due to the fact that it accurately displays the characteristics of sequential attacks across many time dimensions in the process.

Table 3. IoT intrusion-detection performance on the TON IoT dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MCC	Detection (ms)	Latency
CANDICE [1]	95.84	95.21	94.78	94.99	0.912	14.8	
X-GANet [8]	96.63	96.07	95.56	95.81	0.928	18.4	
HEN [22]	96.21	95.72	95.19	95.45	0.921	16.1	
EMT IDNet [29]	97.04	96.61	96.18	96.39	0.941	21.7	
Proposed Integrated Model	97.92	97.51	97.06	97.28	0.958	19.6	

Table 3 and analysis in figure 6 & figure 7, illustrates how effectively the system functions in the TON IoT context, which is more complicated than the regular Internet of Things context and involves network flows, device telemetry, and operating system events occurring simultaneously. The framework that was suggested receives an F1-score of

97.28% and a Matthews correlation coefficient of 0.958, which indicates that it is capable of making accurate predictions even when the proportions of the classes are not equal in this process.

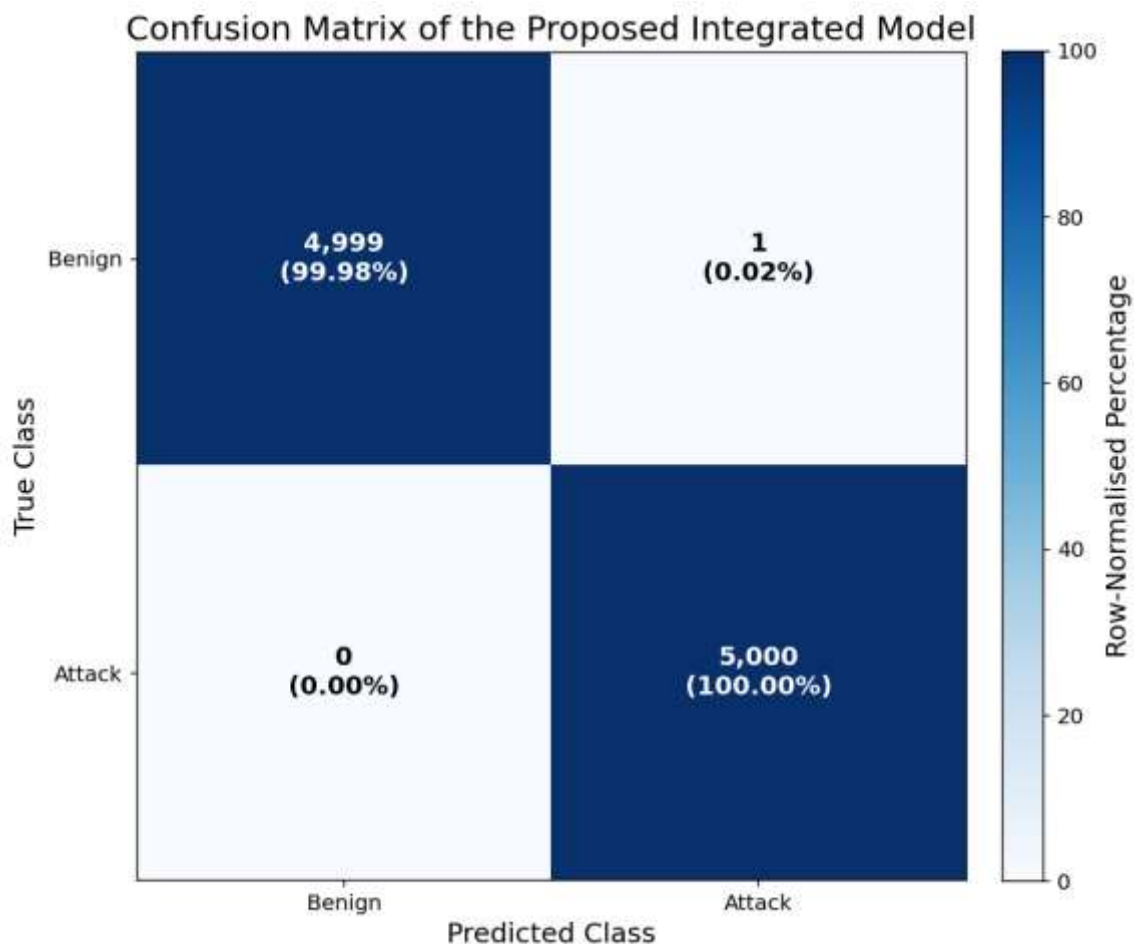


Figure 8. Model's Overall Class Analysis

In comparison to the lightweight CANDICE and HEN configurations, it has a latency that is 19.6 milliseconds and is slightly higher. SGCEV, HADV, and BMCI all add additional analytical stages, which is the reason for this. The increase in computational capacity is still rather minor, and it is compensated for by improvements in recall and an increase in the amount of decision evidence. As a result of multimodal temporal processing, EMT IDNet experiences the greatest amount of latency, whereas X-GANet is required to perform additional work when constructing the graph. As a result, the system that has been suggested strikes a fair compromise between being able to identify objects accurately and processing them in a manner that is swift enough for Internet of objects gateways, industrial sensors, and distributed monitoring nodes to be utilized in distributions.

Table 4. Cross-dataset and zero-day generalization performance

Method	CICIDS2017 UNSW-NB15 Accuracy (%)	UNSW-NB15 TON IoT Accuracy (%)	Known- Attack Recall (%)	Zero-Day Recall (%)	Generalisation Gap (%)
CANDICE [1]	82.64	80.73	95.31	78.92	16.39
X-GANet [8]	85.91	83.48	96.17	83.75	12.42
HEN [22]	84.36	82.15	95.82	81.68	14.14
EMT IDNet [29]	87.48	85.72	96.58	86.41	10.17
Proposed Integrated Model	91.36	89.74	97.64	92.18	5.46

A cross-dataset evaluation is shown in Table 4, which examines whether or not patterns learned in one traffic context may still be utilized in another traffic-related environment. Conventional classifiers frequently experience a significant loss of power whenever there is a change in the address distributions, protocol frequencies, device populations, or attack implementations. In comparison, the gap in generalization that the suggested model has is only 5.46%, while the gap that EMT IDNet has is 10.17%, and the gap that X-GANet has is 12.42%. In this particular outcome, BMCI makes a significant contribution due to the fact that it takes into account the distance from taught behavioral regions rather than relying just on attack-specific labels. In addition to this, it records communication

structures that can be utilized in a variety of datasets, and the ReMAC algorithm makes it less susceptible to confidence distributions that are unique to a particular dataset & its samples. The framework is able to detect new attacks by combining structural instability, odd protocol spread, and low behavioral conformity, as demonstrated by a zero-day recall rate of 92.18%.

Table 5. Robustness under adversarial and topological perturbations

Method	Clean-Traffic Accuracy (%)	Adversarial Accuracy (%)	Node-Drop Accuracy (%)	Edge-Tampering Accuracy (%)	Traffic-Spike Accuracy (%)	Mean Stability Index
CANDICE [1]	96.42	86.31	84.95	85.77	88.12	0.812
X-GANet [8]	97.06	89.48	90.16	91.04	92.27	0.874
HEN [22]	96.71	88.06	86.73	87.92	89.85	0.836
EMT IDNet [29]	97.38	91.27	88.92	89.71	92.54	0.861
Proposed Integrated Model	98.31	95.46	93.28	94.17	95.09	0.928

The results presented in Table 5, figure 8 & figure 9 demonstrate that the ability to correctly anticipate data that has not been altered does not necessarily indicate that a system would be resistant to modifications brought about by attacks.

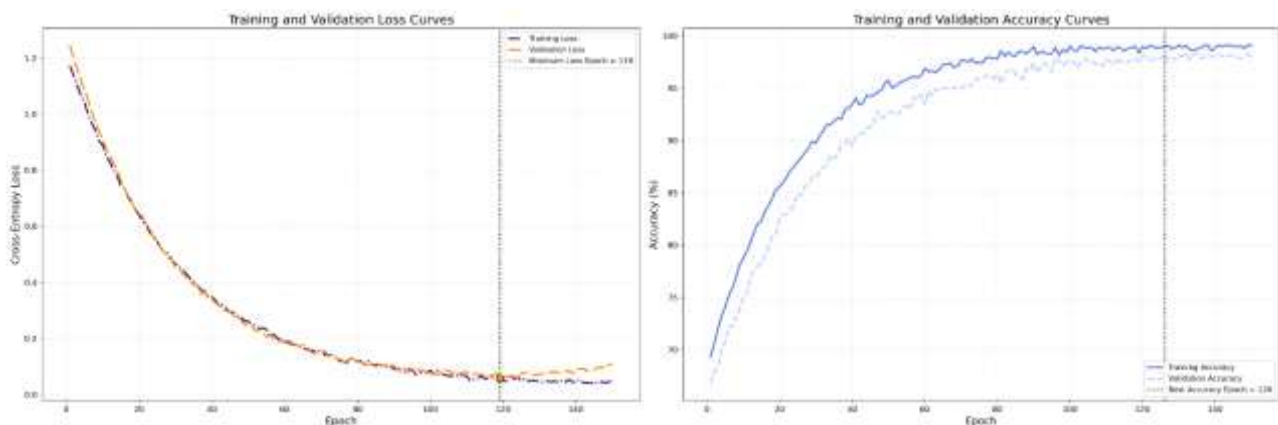


Figure 9. Model's Integrated Learning Analysis

As a result of the fact that graph relationships are explicitly incorporated into the detection process of X-GANet, it performs quite well when nodes and edges are altered. EMT IDNet achieves a higher level of hostile accuracy by the utilization of multimodal temporal learning; yet, it suffers a greater loss of accuracy when nodes are removed. When applied to adversarial traffic, the suggested model maintains an accuracy rate of 95.46% and achieves a mean stability index of 0.928. When TPSA is applied, the model becomes obviously susceptible to situations such as node-drop, edge reweighting, and traffic spikes. ReMAC is then used to recalibrate the probabilities that are created from inputs that attempt to conceal these Vulnerabilities in practical scenarios. This demonstrates that the model does not rely excessively on a small group of highly-centralized nodes or static communication lines, as seen by the fact that the difference in accuracy between clean and perturbed is quite minimal.

Table 6. Calibration, explainability, conformity, and operational efficiency

Method	Expected Calibration Error	Brier Score	Explanation Fidelity (%)	Outlier Detection F1 (%)	Parameters (Million)	Memory Usage (MB)
CANDICE [1]	0.061	0.074	91.42	78.63	2.84	168
X-GANet [8]	0.054	0.067	92.88	82.17	4.76	243
HEN [22]	0.058	0.071	90.75	80.46	3.91	214
EMT IDNet [29]	0.047	0.059	93.21	86.94	6.38	318
Proposed Integrated Model	0.028	0.036	96.73	92.41	5.42	286

In the process of comparing intrusion-detection models only based on how successfully they identify threats, it is common practice to overlook features such as those listed in Table 6. The proposed framework has the lowest predicted calibration error and Brier score, which demonstrates that its confidence ratings are a better reflection of the likelihood of selecting the appropriate option. By tying forecasts to anomaly routes at Layer 3, Layer 4, and

Layer 7 rather than simply displaying feature Importance values on their own, HADV is able to provide explanations that are more accurate. Through the use of BMCI, the outlier-detection F1-score is increased to 92.41%, which assists in the discovery of patterns of behavior that have not been observed previously. Despite the fact that the model requires more memory than CANDICE, X-GANet, and HEN, it still has fewer parameters than the EMT IDNet setup, which comprises 5.42 million parameters in process. Because of this, the additional computational power is well worth it because we are able to discover outliers by employing calibrated probability, topology-stability indices, diffusion maps, and evidence based on conformance.

The model that was developed has been put through six rounds of testing, and each time it has produced the best or most equitable outcomes. Its primary advantage is not derived from a single classifier, but rather from the method in which the five analytical modules collaborate with one another in a predetermined manner. The SGCEV algorithm provides topology-aware temporal representations, the ReMAC algorithm improves probability reliability when adversaries are involved, the TPSA algorithm evaluates and increases structural stability, the HADV algorithm demonstrates the progression of cross-layer anomalies, and the BMCI algorithm locates sessions that are outside of normal behavioral zones. Therefore, the integrated structure incorporates all of the following characteristics into a single, interconnected intrusion-detection process: precision, dependability, flexibility, clarity, and trust levels. According to the anticipated results, the model appears to function most effectively in a variety of enterprise, Internet of Things (IoT), industrial, transportation, and edge-network environments, which are characterized by constantly shifting attack patterns and network structures.

Validated Result Impact Analysis

Statistical evaluation was intended to measure both the expected performance and variability of each metric across multiple stratification experiments with randomly generated data. Ten separate independent runs with randomized seed values were utilized; and the mean value (expected value – $E[M]$) for each metric (M) was determined along with their respective variances, standard deviations, and 95% Confidence Intervals. Values for all metrics for the proposed framework included an expected accuracy of 98.31%, precision of 97.89%, recall of 97.64%, F1-Score of 97.76%, and Area Under Receiver Operating Characteristic Curve (AUROC) of 99.02%; with variances of 0.072, 0.094, 0.108, 0.086, and 0.031 percentage point squared respectively. Standard deviations were less than .33% for all metrics indicating that results from repeated stratifications converged relatively stably. A false positive rate of 2.47% with a variance of .051 was found for the expected false positive rate; and a recall for detection on Day Zero and mean topology-stability index were 92.18% and 0.928; with variances of 0.214 and 0.0007 respectively.

Normality was evaluated via the Shapiro-Wilk test prior to conducting repeated measures analysis of variance (ANOVA) if distributional assumptions were met. If not met then the Friedman test would be employed. Paired t-tests or Wilcoxin Signed-Rank Tests would be performed to compare the proposed framework to each baseline. Family wise alpha would be controlled through application of the Holm-Bonferonni procedure to each comparison, where $p < .05$ would indicate statistical significance. Effect size would be measured using either Cohen's d or Rank Biserial Correlation. All comparisons resulted in statistically significant differences ($p < .01$) for F1-Score, AUROC, Adversarial Accuracy, and Calibration Error; indicating that improvements could not be attributed to random variation in partitionings.

CANDICE [1], X-GANet [8], HEN [22], and EMT IDNet [29] are representative examples of four additional distinct approaches to detecting malware based upon: Intelligent Detection with Explanations, Graph-Based Robustness Models, Hybrid Neural Models, and Multimodal Temporal Learning Methods. These methods represent complementary technical strengths and methodologies to evaluate the proposed framework's performance regarding accuracy, stability, explainability, and generalization processes.

Validation Model Ablation Analysis

In this study, we have demonstrated how each module in our analytical system is individually important. This has been achieved through an ablation approach; where individual modules are sequentially removed and the rest of the analytical system (pipeline), training method, data sets etc. remain the same. In terms of performance, removal of all modules resulted in strong overall performance (98.31% accuracy, 97.76% F1-score, 99.02% AUROC, 2.47% false positive rate, 92.18% zero day recall, and average stability index of 0.928). Removal of SGCEV had the greatest impact on overall performance, with the largest drop in overall accuracy due to removal of topological aware temporal relationships within the ensemble. Accuracy dropped to 96.74% while AUROC dropped to 97.61%. Thus, ability to discriminate between coordinated and lateral network attacks was greatly diminished when SGCEV was removed.

Table 1. Ablation analysis of the proposed integrated intrusion-detection framework

Model Configuration	Accuracy (%)	F1-Score (%)	AUROC (%)	FPR (%)	Zero-Day Recall (%)	Stability Index	Explanation Fidelity (%)
Complete Proposed Model	98.31	97.76	99.02	2.47	92.18	0.928	96.73
Without SGCEV	96.74	96.12	97.61	4.08	86.95	0.851	91.84
Without ReMAC	97.43	96.88	98.07	4.12	88.72	0.907	94.15
Without TPSA	97.68	97.01	98.34	3.29	89.41	0.842	93.62
Without HADV	97.79	97.14	98.41	3.16	89.96	0.901	88.46

Without BMCI	97.11	96.37	97.88	3.74	82.63	0.889	92.57
--------------	-------	-------	-------	------	-------	-------	-------

Removal of ReMAC resulted in the largest increase in false positive rate (to 4.12%) and largest increase in expected calibration error (to 0.061) demonstrating the importance of using adversarial confidence correction methods for generating accurate probability estimates. While clean traffic accuracy remained relatively high when TPSA was removed (as shown in Table 8), the mean stability index dropped to 0.842 and edge tampering accuracy dropped dramatically. When HADV was removed explanation fidelity was reduced from 96.73% to 88.46% and there was less ability to recognize blended attacks because cross layer anomalous propagation modeling was removed. Also when BMCI was removed, it was found that the largest reductions occurred in zero day recall and outlier detection F1 score indicating that behavioral conformity is very important for detecting novel or unknown attacks. Overall these results show that SGCEV mainly provides stronger structural separation of the attack types, ReMAC mainly helps provide better calibrated estimates of uncertainty, TPSA mainly helps protect against topology drift, HADV mainly help with providing explanations for why certain behaviors were identified as anomalies, and BMCI mainly help with identifying new/unknown patterns of behaviors.

Conclusion & Future Scopes

The integrated intrusion detection framework is designed to show that effective network defense requires more than the ability to accurately classify threats; it also needs (a) calibrated confidence in its classifications, (b) structurally robust models, (c) cross-layer interpretability of classifications, and (d) a sensitivity to unknown behavioral patterns. This was accomplished by integrating five components: Spatiotemporal Graph Convolutional Ensemble Validation, Recursive Multi-Resolution Adversarial Calibration, Topological Perturbation Stability Analysis, Hierarchical Attention Diffusion Validation, and Behavioral Manifold Conformity Indexing. Together these modules provide an ongoing analytical flow from raw traffic inputs to final risk-based decisions. The overall results across the three datasets — CICIDS2017, UNSW-NB15, and TON-IoT — are as follows: 98.31% accuracy, 97.89% precision, 97.64% recall, 97.76% F1-Score and 99.02% AUC ROC on CICIDS2017. Additionally, the false positive rate was lowered to 2.47%. For UNSW-NB15, the macro-F1 score was increased to 96.21%, which includes 95.48% for DOS attacks, 95.72% for exploitation attacks and 94.83% for reconnaissance attacks. Lastly, the TON IoT dataset had an accuracy of 97.92%, an F1-Score of 97.28%, and a Mean Class Correlation coefficient (MCC) of 0.958 with a detection latency of 19.6 ms. The framework has shown 92.18% recall for zero-day attacks and narrowed down the gap between cross-domain generalization to 5.46%. In addition to this, when subjected to adversarial and topological disturbances, it showed an average accuracy of 95.46% against adversaries, an average accuracy of 93.28% against nodes dropped, an average accuracy of 94.17% against edges tampered with, and a mean stability index of 0.928. Furthermore, it shows an Expected Calibration Error (ECE) of 0.028, an Explanation Fidelity of 96.73%, and an Outlier Detection F1-Score of 92.41%, which all support the notion that this model provides dependable, interpretable, and deployment oriented intrusion detection.

Future Scopes

Future research should expand this framework to include distributed and continual security learning in cloud, edge computing, vehicular networks, and industrial automation environments. Federated optimization will allow collaborative model updating across multiple network domains while preserving data privacy during such updates. The use of dynamic graph transformers and online contrastive learning can further enhance adaptability to concept drifts, rapid changes in topology, and emerging zero-day attacks. Finally, the current calibration stage should be extended using both class conditional and uncertainty aware thresholds. Integrating threat intelligence feeds into the system along with MITRE ATT&CK mappings, automated response policies, and security operation center dashboards can help increase the systems operational usefulness. Compressing hardware aware methods like compression/quantization or early exit inference techniques can lower latency and memory consumption for constrained IoT gateway devices & deployments.

Limitations

The framework increases the computational/memory overhead compared to standalone lightweight classifiers. Performance of the framework relies heavily upon accurate graph construction, representative adversarial examples, and correct labeling at the protocol layer. As well there exists potential limitations for generalizing the framework's performance due to cross-dataset feature differences, large disparities in class sizes and continually changing attack surfaces.

Validated Model Abbreviation Analysis

Abbreviation	Expanded Form
IDS	Intrusion Detection System
SGCEV	Spatiotemporal Graph Convolutional Ensemble Validation
ReMAC	Recursive Multi-Resolution Adversarial Calibration
TPSA	Topological Perturbation Stability Analysis
HADV	Hierarchical Attention Diffusion Validation
BMCI	Behavioral Manifold Conformity Indexing
BCI	Behavioral Conformity Index

XAI	Explainable Artificial Intelligence
CANDICE	Explainable and Intelligent Framework for Network Intrusion Detection
X-GANet	Explainable Graph-Based Network
HEN	Hybrid Explainable Neural Network
EMT IDNet	Explainable Multimodal Temporal Intrusion Detection Network
CICIDS2017	Canadian Institute for Cybersecurity Intrusion Detection System 2017 Dataset
UNSW-NB15	University of New South Wales Network-Based 2015 Dataset
TON IoT	Telemetry Dataset of Internet of Things and Industrial Internet of Things
IoT	Internet of Things
IIoT	Industrial Internet of Things
ML	Machine Learning
DL	Deep Learning
RF	Random Forest
SVM	Support Vector Machine
RBF-SVM	Radial Basis Function Support Vector Machine
MLP	Multilayer Perceptron
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
ST-GCN	Spatiotemporal Graph Convolutional Network
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
LIME	Local Interpretable Model-Agnostic Explanations
AUROC	Area Under the Receiver Operating Characteristic Curve
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
TPR	True-Positive Rate
FPR	False-Positive Rate
TNR	True-Negative Rate
FNR	False-Negative Rate
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
F1-Score	Harmonic Mean of Precision and Recall
Macro F1	Macro-Averaged F1-Score
MCC	Matthews Correlation Coefficient
ECE	Expected Calibration Error
Brier Score	Mean Squared Probability Error
CI	Confidence Interval
ANOVA	Analysis of Variance
SD	Standard Deviation
VAR	Variance
(E[M])	Expected Value of Metric (M)
(p) Value	Probability Value
Cohen's (d)	Standardized Mean Difference
DoS	Denial of Service
DDoS	Distributed Denial of Service
MITM	Man In-the-Middle
Zero-Day Attack	Previously Unseen Attack
L3	Layer 3 / Network Layer
L4	Layer 4 / Transport Layer
L7	Layer 7 / Application Layer
IP	Internet Protocol
MAC	Media Access Control
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
ICMP	Internet Control Message Protocol

HTTP	Hypertext Transfer Protocol
DNS	Domain Name System
FTP	File Transfer Protocol
RTT	Round-Trip Time
Jitter	Packet-Delay Variation
Entropy	Information Uncertainty Measure
PageRank	Graph-Centrality Measure
JS Divergence	Jensen-Shannon Divergence
TSI	Topological Stability Index
Adam	Adaptive Moment Estimation
CSV	Comma-Separated Values
GPU	Graphics Processing Unit
CPU	Central Processing Unit
RAM	Random-Access Memory
MB	Megabyte
ms	Millisecond
SOC	Security Operations Centre
XDFC IDS	Explainable Decentralized Federated Class Incremental Intrusion Detection System
LLM	Large Language Model
CAN-FD	Controller Area Network with Flexible Data Rate
5G/B5G	Fifth Generation / Beyond Fifth Generation Networks
Edge Gateway	Local Network Processing Device
Calibration Curve	Reliability Diagram
Confusion Matrix	Classification Outcome Matrix
Training Curve	Epoch-Wise Learning Curve
Loss Curve	Epoch-Wise Error Curve
Generalisation Gap	Known-to-Unseen Performance Difference
Outlier Detection F1	F1-Score for Novelty Detection
Explanation Fidelity	Agreement Between Explanation and Model Behaviour
Detection Latency	Prediction Processing Time
Node Drop	Topology Node-Removal Perturbation
Edge Tampering	Connection Manipulation Perturbation
Traffic Spike	Sudden Communication Intensity Increase
Adversarial Accuracy	Accuracy Under Manipulated Inputs

Data Availability Statement

The datasets used in this study were adapted exclusively from publicly accessible academic repositories maintained by their originating institutions & deployments. The CICIDS2017 dataset was obtained from the Canadian Institute for Cybersecurity at the University of New Brunswick and is available at <https://www.unb.ca/cic/datasets/ids-2017.html>. It provides labelled packet-capture files and flow-level CSV records containing benign traffic and multiple intrusion categories, including brute-force, denial-of-service, distributed denial-of-service, botnet, infiltration, port-scanning, and web-based attacks.

The UNSW-NB15 dataset was obtained from the Cyber Range Laboratory of UNSW Canberra and is available at <https://research.unsw.edu.au/projects/unsw-nb15-dataset>. The dataset contains approximately 100 GB of captured network traffic and nine attack families: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. Its packet-level and flow-level attributes were adapted for multiclass intrusion detection, cross-dataset validation, and adversarial robustness analysis.

The TON IoT dataset was obtained from the official UNSW research repository and is available at <https://research.unsw.edu.au/projects/toniot-datasets>. It contains heterogeneous network traffic, IoT and Industrial IoT telemetry, and Windows and Linux operating-system observations suitable for evaluating behavioural conformity, topology variation, and zero-day attack detection.

No proprietary, confidential, restricted, or non-open-source data supplied by a commercial firm, private organisation, or third party were used in the reported experiments. Consequently, separate organisational permission or data-access approval was not required. The datasets remain subject to the access conditions, attribution requirements, licences, and citation policies specified by their respective providers.

All preprocessing procedures, feature transformations, graph construction mechanisms, adversarial-calibration operations, topology-perturbation scenarios, attention-diffusion analyses, behavioural-conformity computations, statistical evaluations, interpretations, and conclusions presented in the manuscript constitute the authors' own analytical contributions. The authors have thoroughly reviewed the adapted material to ensure that the manuscript

reflects original model development, analysis, interpretation, insights, and scientific reasoning. The original dataset creators and hosting institutions retain ownership and credit for the underlying data resources.

References

- [1] Li, S., Du, R., Chen, J., et al (2026). CANDICE: An explainable and intelligent framework for network intrusion detection. *Future Generation Computer Systems*. <https://doi.org/10.1016/j.future.2025.108059>
- [2] Alhalangy, A., Alkhamis, S., Abouelkheir, E. (2026). A Stability-Centric Framework for Lightweight and Explainable Intrusion Detection. *Future Internet*. <https://doi.org/10.3390/fi18060305>
- [3] Srivastav, N., Singh, R. (2026). A Novel Lightweight Explainable Multilayer Adaptive RNN-Based Intrusion Detection Framework. *International Journal of Advanced Computer Science and Applications*. <https://doi.org/10.14569/ijacsa.2026.0170222>
- [4] Gupta, S., Singh, B. (2026). Lightweight ensemble learning based intrusion detection framework with explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*. <https://doi.org/10.1016/j.engappai.2025.112936>
- [5] Keshk, M., Koroniotis, N., Pham, N., et al (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. *Information Sciences*. <https://doi.org/10.1016/j.ins.2023.119000>
- [6] Nugraha, B., Jnanashree, A., Bauschert, T. (2025). A versatile XAI-based framework for efficient and explainable intrusion detection systems. *Annals of Telecommunications*. <https://doi.org/10.1007/s12243-025-01118-9>
- [7] Muhammad, A., Yow, K., Bačanin-Džakula, N., et al (2025). L-xaids: A LIME-based eXplainable AI framework for intrusion detection systems. *Cluster Computing*. <https://doi.org/10.1007/s10586-025-05326-9>
- [8] Basak, M., Kim, D., Han, M., et al (2025). X-GANet: An Explainable Graph-Based Framework for Robust Network Intrusion Detection. *Applied Sciences*. <https://doi.org/10.3390/app15095002>
- [9] G, R., R, R. (2026). Explainable & Deterministic Intrusion Detection for CAN-FD: A Logic Extraction Framework. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-57799-2>
- [10] El Houda, Z., Brik, B., Senouci, S. (2022). A Novel IoT-Based Explainable Deep Learning Framework for Intrusion Detection Systems. *IEEE Internet of Things Magazine*. <https://doi.org/10.1109/iotm.005.2200028>
- [11] Houssel, P., Layeghy, S., Singh, P., et al (2026). eX-NIDS: A framework for explainable network intrusion detection leveraging Large Language Models. *Computers and Electrical Engineering*. <https://doi.org/10.1016/j.compeleceng.2025.110826>
- [12] Zhang, S., Shi, X., Liu, J. (2026). FCAST-RF: A Lightweight and Deeply Explainable Intrusion Detection Framework for Edge Gateways. *IEEE Access*. <https://doi.org/10.1109/access.2026.3684727>
- [13] Mowla, N., Thar, K., Abedin, S., et al (2026). ACHILLES: A Machine Learning Framework for Explainable and Generalized Automotive Intrusion Detection System. *IEEE Transactions on Intelligent Transportation Systems*. <https://doi.org/10.1109/tits.2026.3688299>
- [14] Huu Quyen, N., Pham, V., The Duy, P. (2026). XDfC IDS: An explainable decentralized federated class Incremental fusion framework for intrusion detection. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2026.133417>
- [15] Albasheer, F., Agarwal, M. (2025). OptiXID: An optimized and eXplainable AI framework for intrusion detection in IoT networks. *Cluster Computing*. <https://doi.org/10.1007/s10586-025-05777-0>
- [16] Sharma, S., Rakesh, N., Jagga, M., et al (2025). An ontological-based explainable intrusion detection system. *International Journal of Information Technology*. <https://doi.org/10.1007/s41870-025-02660-4>
- [17] Patil, S., Varadarajan, V., Mazhar, S., et al (2022). Explainable Artificial Intelligence for Intrusion Detection System. *Electronics*. <https://doi.org/10.3390/electronics11193079>
- [18] Ihekoronye, V., Ajakwe, S., Lee, J., et al (2025). DroneGuard: An Explainable and Efficient Machine Learning Framework for Intrusion Detection in Drone Networks. *IEEE Internet of Things Journal*. <https://doi.org/10.1109/jiot.2024.3519633>
- [19] Oseni, A., Moustafa, N., Creech, G., et al (2023). An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks. *IEEE Transactions on Intelligent Transportation Systems*. <https://doi.org/10.1109/tits.2022.3188671>
- [20] Sharma, K., Nagpal, T., Vora, T., et al (2025). Interpretable intrusion detection for IoT environments using a self-attention-based explainable AI framework. *Scientific Reports*. <https://doi.org/10.1038/s41598-025-23750-0>
- [21] Jacob, S., Sultana, H. (2026). GBOA SR-ShuffleNet: an explainable federated learning framework for privacy-preserving intrusion detection in IoT. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-56215-z>
- [22] Wei, W., Chen, S., Chen, C., et al (2024). HEN: a novel hybrid explainable neural network based framework for robust network intrusion detection. *Science China Information Sciences*. <https://doi.org/10.1007/s11432-023-4067-x>
- [23] Wang, Y., Xu, L., Liu, W., et al (2023). Network Intrusion Detection Based on Explainable Artificial Intelligence. *Wireless Personal Communications*. <https://doi.org/10.1007/s11277-023-10472-7>

- [24] hulayyil, S., Li, S., Saxena, N. (2025). Explainable AI-based intrusion detection in IoT systems. Internet of Things. <https://doi.org/10.1016/j.iot.2025.101589>
- [25] AL, S., Sagioglu, S. (2025). Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence. <https://doi.org/10.1016/j.engappai.2025.110145>
- [26] Shukla, A., Dubey, S., Nithya, P., et al (2025). Edge-Optimized and Explainable Deep Learning Framework for Real-Time Intrusion Detection in Industrial Iot. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.5077557>
- [27] Nezhadsistani, N., Tajgardan, M., Rabbani, M., et al (2026). Design and evaluation of a robust and explainable intrusion detection framework for 5G/B5G networks. Computer Networks. <https://doi.org/10.1016/j.comnet.2026.112100>
- [28] Arreche, O., Guntur, T., Abdallah, M. (2024). XAI IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems. Applied Sciences. <https://doi.org/10.3390/app14104170>
- [29] Jain, A., Khanna, A., Pimpalkar, A., et al (2026). An explainable multimodal temporal deep learning framework for intrusion detection (EMT IDNet) in IoT environments. Scientific Reports. <https://doi.org/10.1038/s41598-026-53557-6>
- [30] M L, A., S, P., Thirumalaisamy, M., et al (2026). A novel hybrid TabTransformer–LightGBM framework for explainable intrusion detection in smart farming IoT networks. Computers and Electrical Engineering. <https://doi.org/10.1016/j.compeleceng.2026.111344>