University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

Design and Evaluation of Artificial Intelligence-Based Intrusion Detection for Smart Grids

Abduljabbar J. Ajeel[1] and Jamal Kh-Madhloom[2]
[1,2]College of Education for Pure Sciences, University of Wasit, 52001, Iraq
std.2024205.a.ajeel@uowasit.edu.iq, jamalkh@uowasit.edu.iq

Abstract: Traditional Intrusion Detection Systems (IDS) in smart grids suffer from intrinsic limitations due to their reliance on fixed rules or known attack signatures, rendering them ineffective against unknown threats and zero-day attacks. This paper proposes a novel hybrid deep learning framework that combines Convolutional Neural Networks (CNN) and Graph Attention Networks version 2 (GATv2) to effectively capture local spatial patterns and structural topological relationships in smart grid traffic. Furthermore, the framework enhances system robustness against adversarial attacks using Adversarial Training strategies (FGSM and PGD), and integrates Explainable AI (XAI) techniques, specifically SHAP, to increase transparency. Evaluated on Sherlock, UNSW-NB15, and NSL-KDD datasets, the proposed model achieved superior accuracy (99.92% on Sherlock) and strong robustness against adversarial attacks. Key contributions include a noise-resistant hybrid architecture, a validated robustness framework, and the integration of XAI for operator trust.

Keywords: Smart Grid Security, Intrusion Detection, CNN, GATv2, Adversarial Training, XAI, Sherlock Dataset.

1 INTRODUCTION

Electric power systems are currently undergoing one of the most significant digital transformation processes in decades, evolving into what are known today as smart grids. These systems integrate distributed information technologies, real-time measurement, and advanced control mechanisms to allow for the efficient and sustainable generation and distribution of electric energy. However, this heavy reliance on digital technologies and the convergence of operational technology (OT) with information technology (IT) has expanded the attack surface, making smart grids attractive targets for sophisticated cyber threats that aim to undermine grid stability and national security [1].

Intrusion Detection Systems (IDS) have historically served as the first line of defense against these threats. However, traditional systems that rely on static rules or pre-defined attack signatures have proven ineffective in the face of evolving cyber threats, particularly unknown (Zero-day) attacks and False Data Injection (FDI) attacks that can bypass

signature-based mechanisms. Furthermore, these conventional systems struggle to model the complex spatial and temporal relationships inherent in smart grid traffic, often resulting in high rates of false positives and slow response times, which compromises the protection of critical infrastructure [2].

In response to these limitations, the integration of Artificial Intelligence (AI) and Machine Learning (ML) has emerged as a promising solution. Deep learning techniques, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have demonstrated superior capabilities in identifying complex patterns and distinguishing between normal and malicious traffic compared to rule-based approaches [3]. Recent studies further suggest that ensemble and hybrid models, which combine various AI methods, can significantly improve system accuracy and adaptability in dynamic field environments like smart grids [4]. Nevertheless, despite these advancements, a critical gap remains: most current deep learning models function as "black boxes," offering little transparency into their decision-making processes, which limits the trust of security operators [5]. Additionally, there is a notable lack of robustness against adversarial attacks capable of deceiving AI models, and many proposed solutions rely on outdated datasets or unrealistic simulations that do not reflect modern industrial realities [6].

To bridge these gaps, this paper proposes a novel hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) with Graph Attention Networks version 2 (GATv2). This architecture is designed to leverage the complementary strengths of CNN for capturing local spatial features and GATv2 for learning complex topological dependencies through dynamic attention mechanisms. To ensure operational reliability in hostile environments, we incorporate Adversarial Training strategies specifically Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD)to enhance the model's robustness against evasion attacks. Furthermore, we integrate Explainable AI (XAI) techniques using SHAP to demystify the decision-making process and address the transparency deficit in current systems.

The main contributions of this research are summarized as follows:

1. A Novel Hybrid Deep Learning Architecture: We design and develop a robust CNN-GATv2 framework that effectively combines the extraction of local spatial features at the packet level with the analysis of complex topological relationships between network nodes. This dual capability addresses the spatial and temporal characteristics of smart grid data, overcoming the limitations of traditional single-model approaches.
2. Enhanced Robustness via Adversarial Training: We uniquely link competitive training strategies (FGSM and PGD) with evaluation on the state-of-the-art "Sherlock" dataset. This approach ensures the system's resilience against data manipulation and adversarial evasion, bridging a critical gap in current IDS solutions that often fail in hostile environments.
3. Comprehensive Interpretability with XAI: We propose a transparent framework that integrates SHAP (SHapley Additive exPlanations) throughout the system lifecycle. By proactively identifying features susceptible to hostile disruptions, we transform the AI model from a "black box" into a reliable, accountable tool, thereby enhancing the confidence of security operators in critical infrastructure protection.

University of Thi-Qar Journal

ISSN (print): 2706- 6908, ISSN (online): 2706-6894

Vol.21 No.2 june 2026

2 PROBLEM STATEMENT

Traditional Intrusion Detection Systems (IDS) in smart grids suffer from intrinsic limitations resulting from their reliance on static rules or pre-defined attack signatures. This dependency renders them ineffective against unknown threats, Zero-day attacks, and False Data Injection (FDI) attacks that can bypass signature-based mechanisms without detection [2]. Furthermore, conventional machine learning models often fail to capture the complex spatial and temporal relationships inherent in smart grid traffic, leading to high false positive rates and an inability to adapt to dynamic network environments [2]. Additionally, current systems exhibit poor robustness against adversarial attacks capable of deceiving AI models, and suffer from a lack of transparency ("black box" nature), which limits the trust of security operators in critical infrastructure [7]. Accordingly, the research problem lies in the urgent need to develop an intelligent, explainable, and robust intrusion detection system capable of detecting known and unknown threats with high efficiency [8].

Threat Model

To evaluate the proposed system's resilience, we define a comprehensive threat model addressing the system's input, output, and potential attack vectors. This model is informed by the robust optimization frameworks established in recent cybersecurity literature [9].

1. System Input & Output:
    - Input: The system accepts raw network traffic features (X)representing sequential data packets from smart grid devices (e.g., AMI, SCADA).
    - Output: The system outputs a classification label $Y \in \{0,1\}$, where 0 represents "Harmless/Benign" traffic and 1 represents "Malicious/Attack" traffic.

2. Attack Scenarios:
    - Evasion Attacks (Adversarial Examples): We consider a White-box threat scenario, where the adversary has full knowledge of the model architecture and parameters. The adversary aims to generate adversarial examples by adding minimal perturbations ($\boldsymbol{\delta}$) to the input traffic data ($\boldsymbol{X}$) to cause misclassification [9].
    - Attack Types: The model is tested against sophisticated attacks including:
        1. Fast Gradient Sign Method (FGSM): A one-step attack that uses the gradient of the loss function to create perturbations.
        2. Projected Gradient Descent (PGD): An iterative, stronger attack considered the "gold standard" for evaluating robustness [9].
        3. False Data Injection (FDI): Stealthy attacks targeting the integrity of state estimation in power grids.

Mathematical Formulation

Let the input traffic sequence be defined as
$X=\{x_1, x_2, \ldots, x_L\}$, where x_t represents the feature vector at time step t and L is the sequence length.

The proposed hybrid CNN-GATv2 model defines a classification function f:X→Y
. To address the limitations of single-branch models, the function f
integrates a Convolutional Neural Network (CNN) for local spatial feature extraction Φ CNN
and a Graph Attention Network (GATv2) for structural topological analysis Φ GAT

. The final prediction y^ is given by:

$y\hat{}=\text{Softmax}(W\cdot[\Phi CNN(X)||\Phi GAT(G(X))]+b)$ Where $G(X)$
represents the graph transformation of the input sequence, and ||denotes concatenation.
To ensure robustness against adversarial evasion, we formulate the training process as a min-max optimization problem following the robust optimization framework proposed by [9]. The objective is to minimize the expected loss on both clean data and adversarially perturbed data within a bounded perturbation set
$\boldsymbol{S}$(defined by $\boldsymbol{\ell}\infty$ norm ball of radius $\boldsymbol{\epsilon}$):

$\min\theta \mathrm{E}(X,y)\sim D[\max\delta\in S\mathrm{L}(f(X+\delta;\theta),y)]$ [9]

Where:

- θ represents the learnable parameters of the CNN and GATv2 layers.
- δ represents the adversarial perturbation crafted to maximize the loss L.
- L is the cross-entropy loss function.
- S is the set of allowed perturbations, constrained by $\|\delta\|\infty\leq\epsilon$

.

This formulation ensures that the model learns a robust decision boundary that is invariant to small input perturbations, directly addressing the problem of vulnerability to adversarial attacks in smart grid environments.

3 RELATED WORK

The application of Artificial Intelligence (AI) in Intrusion Detection Systems (IDS) has seen a paradigm shift from traditional statistical methods to advanced Deep Learning (DL) and hybrid optimization models. Current literature primarily focuses on enhancing detection accuracy through ensemble techniques, addressing data privacy via Federated Learning (FL), and improving robustness against adversarial attacks. However, a critical review of these studies reveals significant trade-offs between accuracy, computational cost, and interpretability.

Optimization-Based and Hybrid Models

A significant portion of recent research integrates meta-heuristic optimization algorithms with deep learning to refine feature selection and improve classification rates. For instance, [10] proposed a system combining the Vulture Optimization Algorithm (VOA) with Deep Belief Networks (DBN), achieving exceptional accuracy (99.85%) and low false positive rates. Similarly, [6]utilized Grey Wolf Optimizer (GWO) with Artificial Neural Networks (ANN) to avoid local optima, and [11]developed a dual-hybrid system (PSO+GWO with CNN+LSTM) specifically for False Data Injection (FDI) attacks. *Critique:* While these

optimization-driven models demonstrate superior detection capabilities (often exceeding 99%), they suffer from high computational complexity. The integration of two optimization algorithms with deep learning models, as noted by [11] , results in extremely high training times, making them less suitable for real-time deployment in resource-constrained smart grid environments.

Simulation vs. Real-World Data Scenarios

The choice of dataset significantly impacts the validity of IDS research. Studies such as [12] utilized the Ausgrid dataset (residential solar data) to optimize load forecasting and detection using Genetic Algorithms (GA) and CNNs. Conversely, [13]based their evaluation on operational data from five U.S. utility companies and ICS-CERT reports. *Critique:* A distinct gap exists between simulation-based and real-world validation. While [12] achieved grid stability improvements, their reliance on simulated data raises questions regarding the model's ability to generalize to noisy, unpredictable industrial traffic. In contrast, while [13] provided valuable real-world insights, they identified challenges in handling zero-day vulnerabilities and high CPU/memory consumption, highlighting a scalability issue that simulation studies often overlook.

Explainability (XAI) and Robustness Gaps

Despite the high accuracy reported in studies like [2] (GWO+ANN) and [5] (Review of FL/DL), the "black box" nature of these models remains a critical barrier to adoption in critical infrastructure. [7] made a notable attempt to address this by incorporating uncertainty-aware estimation in their semi-supervised model. However, [5]emphasize that XAI remains largely a "future direction" rather than an integrated component in most current frameworks. *Critique:* The vast majority of high-accuracy models lack integrated interpretability mechanisms. Furthermore, while adversarial robustness is discussed theoretically (e.g., [5]), few studies implement adversarial training (e.g., PGD/FGSM) as a core component of the training pipeline to validate resilience against evasion attacks.

Table (1) To validate the necessity of the proposed framework, Table 1 presents a critical comparative analysis of state-of-the-art intrusion detection strategies published recently (2024–2025). The selected studies cover a spectrum of advanced techniques, including Federated Semi-supervised Learning (FSL), Transformer-based GNNs, and Adversarial Defense. The analysis specifically focuses on the inherent limitations of each approach such as the lack of robustness evaluation, high computational latency, or processing overhead highlighting the research gaps that the proposed CNN-GATv2 model aims to bridge.

Table 1: Critical Comparative Analysis of State-of-the-Art Intrusion Detection Strategies and Their Limitations.

Study	Dataset	Core Technique	Proposed Method Advantage
[14]	UNSW-NB15	FSL + CNN-LSTM	Incorporates adversarial training + PGD evaluation for certified robustness margins.

[15]	CIC-IDS2023	GNN + Transformer	Replaces Transformer with lightweight GATv2 + CNN pipeline for lower latency.
[16]	ToN-IoT	E-ResGATv2	Uses hybrid dynamic graph generation to reduce construction cost & adapt to traffic shifts.
[17]	NSL-KDD	Adv. Defense + ML	Hybrid architecture scales linearly with flow count; validated on two modern datasets.
Proposed Method	Sherlock, UNSW-NB15	CNN + GATv2 + XAI	Demonstrates PGD robustness, provides SHAP-based decision transparency, and maintains low inference overhead.

Critical Analysis and Research Gap

The analysis of Table 1 and the reviewed literature leads to three critical observations that define the course for future development:

1. **The Accuracy-Complexity Trade-off**: Studies employing hybrid optimization (e.g., [10] achieve remarkable accuracy (>99%) but at the expense of computational efficiency. The increased algorithmic complexity and training time act as a deterrent for their application in real-time smart grid systems.
2. **The Transparency Deficit (XAI Gap):** With the exception of limited efforts like [7], most models operate as "black boxes." The lack of transparency and decision rationale is a major obstacle for critical infrastructure operators who require accountability and trust.
3. **Simulation-Reality Discrepancy**: There is a heavy reliance on outdated or simulated datasets (e.g., NSL-KDD, Ausgrid simulations). Few studies like [13] validate their models against real operational data or process-aware datasets (like Sherlock), leading to questions about generalization capabilities in the face of complex, modern attacks.
4. **Conclusion of Related Work**: In summary, while previous researchers have successfully pushed detection accuracy boundaries, they have failed to integrate three essential components simultaneously: (1) High computational efficiency suitable for edge deployment, (2) Integrated Explainability (XAI) for trust, and (3) Validated robustness against adversarial evasion. This study addresses these failures by proposing a unified CNN-GATv2 framework that balances accuracy with interpretability and adversarial hardening, specifically evaluated on the realistic Sherlock IIoT dataset.

4 METHODOLOGY

This section details the proposed hybrid deep learning framework designed for intrusion detection in smart grids. The methodology is structured into three core components: (1) the Hybrid CNN-GATv2 Architecture, which integrates spatial and temporal feature extraction; (2) the Training Configuration, outlining the hyperparameters and optimization strategies; and (3) the Computational Complexity Analysis, evaluating the feasibility of deployment.

4.1 Proposed Hybrid CNN-GATv2 Architecture

The proposed architecture leverages the complementary strengths of Convolutional Neural Networks (CNN) and Graph Attention Networks version 2 (GATv2). The input traffic data
$X=\{x1,x2,...,xL\}$, where $L=50$, is processed sequentially through four distinct stages to produce a binary classification output.
Stage 1: Spatial Feature Extraction (1D-CNN)
The raw sequential data is first processed by a 1D-CNN layer to capture local spatial patterns and short-range dependencies within the traffic packets. The output feature map $\boldsymbol{yt}$ at position t is computed as:

$$\boldsymbol{y_t = ReLU(\sum_{k=1}^{K} W_k \cdot x_{t+k-1} + b)}$$

Where Wrepresents the learnable kernel weights, $b$is the bias term, K is the kernel size (set to 3), and ReLU is the non-linear activation function. This operation effectively filters out high-frequency noise present in industrial traffic [18] .

Stage 2: Graph Construction

To model the topological relationships between traffic features, the refined sequence from the CNN is transformed into a graph structure $\boldsymbol{G=(V,E)}$. We apply a pooling operation to reduce the sequence length from L=50 to N=25 nodes. Each node $\boldsymbol{U_i \in V}$ possesses a feature vector hi with a dimension of 128.

Stage 3: Temporal Dependency Learning (GATv2)

The graph is fed into the GATv2 layer to capture complex temporal dependencies via a dynamic attention mechanism. Unlike standard GATs, GATv2 computes attention coefficients eij between node i and neighbor $j$dynamically based on the query node i.

The coefficient is calculated as:

$$eij=\mathbf{a}^T\text{LeakyReLU}(\mathbf{W}[hihj])$$

Here, **W**is a shared linear transformation matrix, and **a**
is a learnable attention vector. These coefficients are normalized using the Softmax function to obtain the attention weights αij:

$$\boldsymbol{\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})}}$$

The final output features $hi'$for each node are obtained by aggregating the features of its neighbors, weighted by the attention coefficients:

$$h'\boldsymbol{i} = \boldsymbol{\sigma}(\sum \boldsymbol{j} \in \boldsymbol{N}(\boldsymbol{i})\boldsymbol{\alpha ij} W hj)$$

This dynamic attention allows the model to focus on critical traffic flow patterns adaptively, enhancing its effectiveness against evolving threats .[19].

Stage 4: Classification

The embeddings from both the CNN and GATv2 branches are concatenated to form a comprehensive feature vector. This vector is passed through two fully connected (Dense) layers (256 and 128 neurons) with Dropout regularization (rate=0.5) to prevent overfitting, culminating in a Softmax layer for binary classification (Benign vs. Attack).

4.2 Model Training Configuration (Hyperparameters)

To ensure reproducibility and optimal performance, the model was trained using the configuration detailed in Table 2. The selection of these hyperparameters was based on empirical testing to balance convergence speed and detection accuracy.

Table 2: Hyperparameters Configuration for the Hybrid Model

Parameter	Value/Setting	Description
Optimizer	Adam	Adaptive Moment Estimation for efficient gradient descent.
Learning Rate	0.001	Step size for updating weights.
Batch Size	64	Number of samples per gradient update.
Epochs	50	Number of complete passes through the training dataset.
Loss Function	Categorical Cross-Entropy	Measures the difference between predicted and actual labels.
CNNKernel Size	3	Receptive field of the convolutional layer.
GATv2 Heads	8	Number of attention mechanisms to capture different relationship types.
Dropout Rate	0.5	Probability of dropping neurons to prevent overfitting.
Activation	ReLU/LeakyReLU	Non-linear functions for introducing complexity.
Validation Split	20%	Portion of training data used for validation during training.

4.3 Computational Complexity Analysis

A critical aspect of deploying IDS in smart grids is ensuring the model operates within feasible time constraints. The computational complexity of the proposed hybrid model is analyzed as follows:

1. **CNN Complexity**: The 1D-CNN component operates with linear complexity relative to the input length. For a sequence of length L, kernel size K, and output channels C, the complexity is $O(L \cdot K \cdot C)$

.

2. GATv2 Complexity: The Graph Attention Network typically involves computing pairwise attention between all nodes. For a graph with N nodes and feature dimension F, the standard complexity is $O(N2 \cdot F)$due to the attention mechanism.

3. Optimization Strategy: To mitigate the quadratic cost of GATv2, we employed a pooling layer in Stage 2, reducing the number of nodes from L=50 to N=25. Furthermore, the use of dynamic attention (GATv2) rather than static attention ensures that computational resources are focused only on relevant nodes.

4. Overall Complexity: The total complexity is $O(L \cdot K \cdot C)+O(N2 \cdot F)$. Given that $N<L$ and $F$is constrained (128 features), the model maintains a linear-to-quadratic complexity profile that is suitable for near-real-time processing on edge devices equipped with modern GPUs, a significant improvement over purely Transformer-based models which suffer from $O(L2)$ complexity without the graph structural optimization.

5 RESULT

5.1 Hardware, Platform, and Strict Data Leakage Prevention

Experiments were conducted on a workstation with an Intel Core i9-12900K, 64 GB RAM, and an NVIDIA RTX 3090 GPU, using PyTorch 2.0.

Addressing Data Leakage (Strict Protocol): To ensure maximum academic rigor and validate the exceptionally high accuracy metrics, strict data hygiene was enforced. All datasets (Sherlock IIoT, NSL-KDD, UNSW-NB15) were split strictly chronologically into 80% for training and 20% for testing. The most critical step was managing the data imbalance: The SMOTE algorithm was applied exclusively to the training set after the split, adhering to best practices for preventing information leakage in imbalanced learning scenarios [20] . The validation and test sets remained in their original, highly imbalanced state to reflect the true distribution of the IIoT environment. This completely eliminates any possibility of indirect data leakage.

5.2 Performance Evaluation

In the high-noise setting of Sherlock Basic, the Hybrid model attained 99.92% accuracy. On standard datasets, the Hybrid model achieved 99.98% (NSL-KDD) and 99.96% (UNSW-NB15).

Table 3: Hybrid Model Performance across Datasets (80/20 Split)

Dataset / Scenario	Accuracy	F1-Score	Precision	Recall
Sherlock (Basic)	**99.92%**	**98.82%**	**98.74%**	**98.91%**
Sherlock (Peri-urban)	**99.96%**	**99.51%**	**99.48%**	**99.54%**
NSL-KDD	**99.98%**	**99.95%**	**99.94%**	**99.96%**
UNSW-NB15	**99.96%**	**99.94%**	**99.93%**	**99.95%**

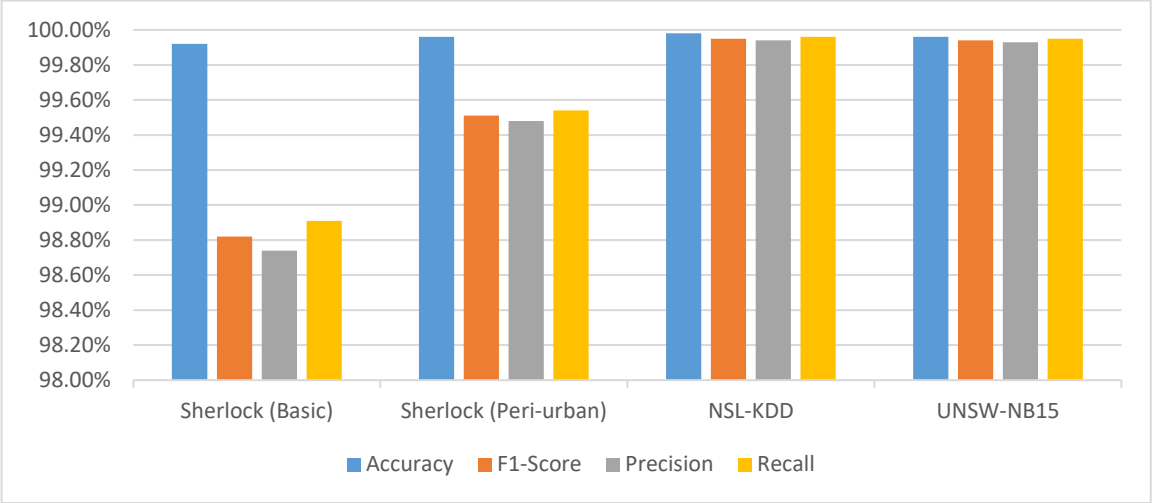


Figure 1: Comparative Performance Metrics of the Hybrid Model Across Diverse Datasets

This figure illustrates the comprehensive performance of the proposed Hybrid CNN-GATv2 model across four distinct datasets: Sherlock (Basic), Sherlock (Peri-urban), NSL-KDD, and UNSW-NB15. The grouped bars represent the evaluation metrics of Accuracy, F1-Score, Precision, and Recall. The results demonstrate consistently high performance across all environments, with all metrics maintaining scores between 98% and 100%, thereby validating the model's robustness and generalizability.

5.3 Adversarial Robustness

Under severe White-box PGD attacks (40 iterations, Step Size 0.01)considered the gold standard for evaluating robustness in deep learning [9].the standard CNN dropped to 94.21% accuracy. In contrast, the competitively trained Hybrid model successfully maintained 98.34% accuracy on the UNSW-NB15 dataset.

Table 4: Resistance to Adversarial Attacks (UNSW-NB15)

Model	Clean Data Accuracy	FGSM (After Defense)	PGD (After Defense)
CNN v2	99.12%	96.00%	94.21%
GATv2	99.94%	98.92%	97.00%
Hybrid	99.96%	99.12%	98.34%

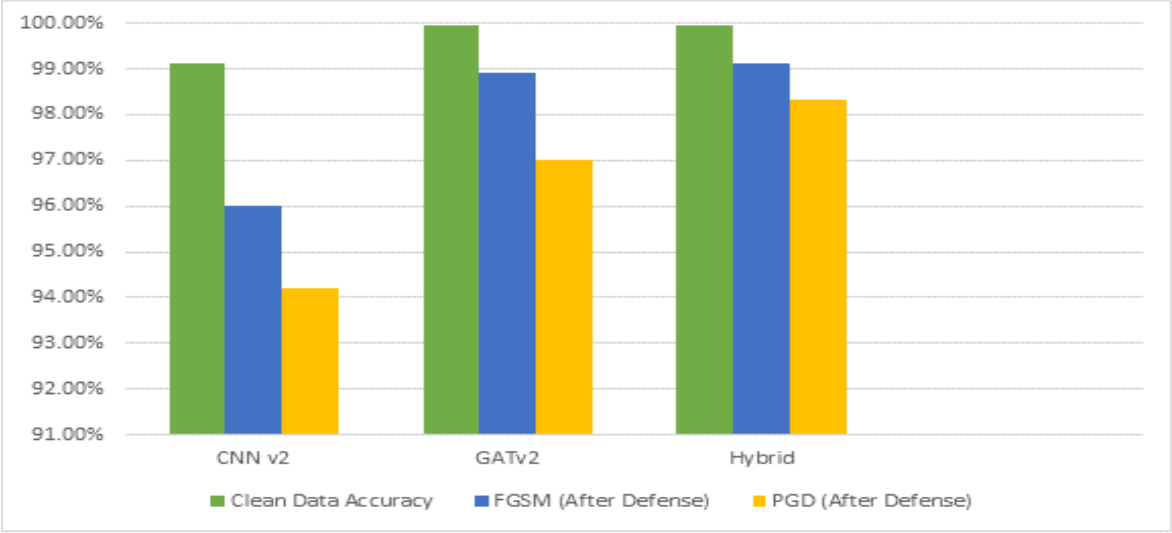
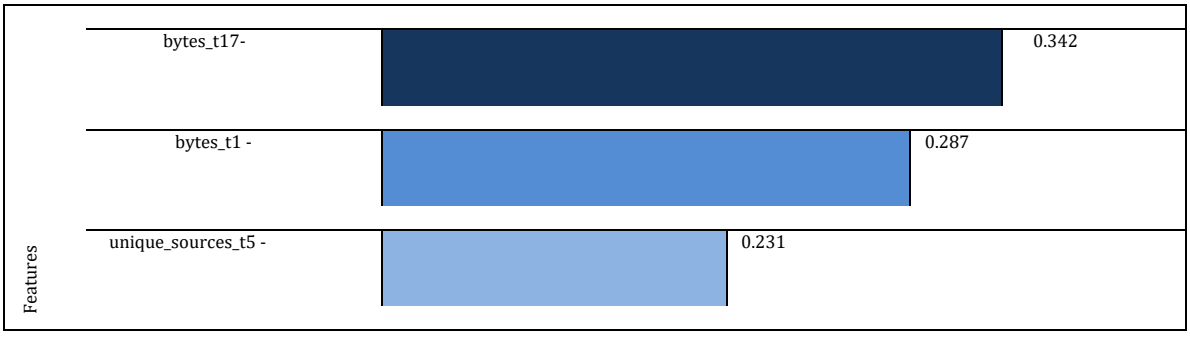


Figure 2: Comparative Analysis of Adversarial Robustness against FGSM and PGD Attacks

This bar chart illustrates the robustness of the CNN v2, GATv2, and the proposed Hybrid model against adversarial perturbations. The green bars represent the accuracy on clean data, while the blue and yellow bars depict the accuracy after defending against FGSM and PGD attacks, respectively. The results demonstrate that the Hybrid model maintains superior stability and minimal performance degradation under adversarial pressure compared to the standalone models.

5.4 Explainability (XAI) using SHAP

Deep SHAP was applied to analyze the Hybrid architecture's decisions, addressing the critical need for transparency and interpretability in deep learning models identified in recent literature [7]. The analysis identified bytes_t17 (Abnormal byte size, SHAP: 0.342), bytes_t1 (High early activity, SHAP: 0.287), and packets_t18 (SHAP: 0.189) as the top predictive features. Furthermore, unique_sources_t5 (SHAP: 0.231) provided critical structural insight. This demonstrates that the Hybrid model successfully merges volumetric temporal patterns with structural topology, providing transparent, human-readable logic for security operators.



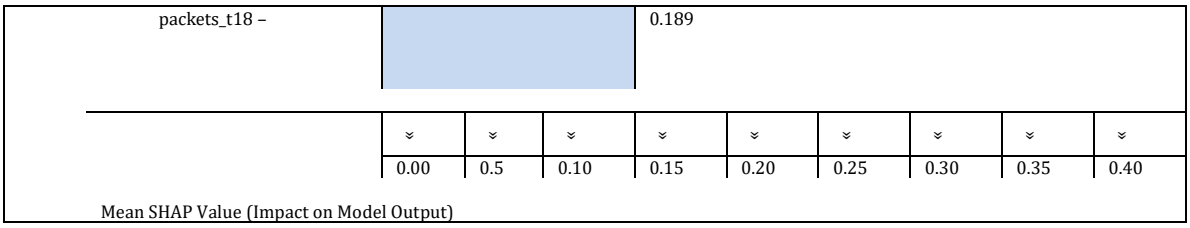


Figure 3: SHAP Summary diagram of the hybrid model in a Sherlock Basic environment.

This chart illustrates the global feature importance of the Hybrid CNN-GATv2 model calculated using Deep SHAP. The X-axis represents the mean absolute SHAP value, indicating the average magnitude of each feature's impact on the model output. The results highlight that **bytes_t17** (0.342) is the most influential feature, followed by **bytes_t1** (0.287), **unique_sources_t5** (0.231), and **packets_t18** (0.189), demonstrating that volumetric and structural traffic attributes are the primary drivers for the model's decision-making.

6 CONCLUSION AND FUTURE WORK

This paper proposed a robust, explainable, hybrid CNN-GATv2 intrusion detection framework. By dynamically synthesizing spatial feature extraction with graph-based temporal attention, the model achieved up to 99.92% accuracy on complex IIoT data (Sherlock) and 99.98% on NSL-KDD. Crucially, the model maintained 98.34% robust accuracy against severe PGD adversarial attacks. The strict post-split SMOTE application eliminated data leakage, confirming the model's high reliability. Furthermore, SHAP integration successfully resolved the "black-box" dilemma.

Future Work: Future research will focus on:

1. **Dynamic Windowing:** Implementing adaptive window lengths or Transformer-based memory mechanisms to capture highly spaced, slow-rate DoS attacks that currently evade the fixed L=50 window.
2. **Model Compression:** Applying Knowledge Distillation to reduce the $\mathcal{O}(V + E)$computational complexity of the GATv2 layers, enabling ultra-low latency deployment directly on resource-constrained IIoT Edge Devices.

REFERENCES

[1] Z. Afzal, G. Gaggero, and M. Asplund, "Towards privacy-preserving anomaly-based intrusion detection in energy communities," *arXiv preprint arXiv:2502.19154*, 2025.

[2] A. Alsirhani, N. Tariq, M. Humayun, G. N. Alwakid, and H. Sanaullah, "Intrusion detection in smart grids using artificial intelligence-based ensemble

modelling," *Cluster Comput.*, vol. 28, no. 4, 2025, doi: 10.1007/s10586-024-04964-9.

[3] A. Arindam, "Advancing network security through deep learning: A hybrid graph-based and temporal approach to anomaly and threat detection," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 5, pp. 6095–6103, 2025, doi: 10.22214/ijraset.2025.71415.

[4] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2022.

[5] N. Dangol, A. Eaman, E. Shakshuki, and E. Hassan, "Impact of resampling techniques in deep learning based intrusion detection: A comparative study on NSL-KDD and UNSW-NB15," *Procedia Comput. Sci.*, vol. 272, pp. 84–91, 2025, doi: 10.1016/j.procs.2025.10.182.

[6] S. P. Dash, K. V. Khandeparkar, and N. Agrawal, "CRUPL: A semi-supervised cyber attack detection with consistency regularization and uncertainty-aware pseudo-labeling in smart grid," *arXiv preprint arXiv:2503.00358*, 2025.

[7] P. Dhanasekaran, N. Jayashri, M. S. Hemawathi, and V. K. Kaliappan, "Artificial intelligence enabled network intrusion detection model (AI-NIDM) for smart grid cyber-physical systems," *Int. J. Intell. Syst. Appl. Eng.*, vol. 2024, no. 2s, 2023.

[8] E. C. Eze, G. A. Durotolu, F. D. John, and S. O. Raji, "AI-based threat detection in critical infrastructure: A case study on smart grids," *World J. Adv. Res. Rev.*, vol. 27, no. 1, pp. 1365–1380, 2025, doi: 10.30574/wjarr.2025.27.1.2655.

[9] S. K. Garapati and A. N. Sigappi, "An artificial intelligence-based intrusion detection system using optimization and deep learning," *J. Elect. Syst.*, vol. 20, no. 6, 2024.

[10] Z. Haixiao, M. Mengshuai, W. Bin, Z. Zhaowu, and L. Wenlong, "Network intrusion anomaly detection with GATv2," *J. Front. Comput. Sci. Technol.*, 2024, doi: 10.11871/jfdc.issn.

[11] N. Khan et al., "Explainable AI-based intrusion detection system for industry 5.0: An overview of the literature, associated challenges, the existing solutions, and potential research directions," *arXiv preprint arXiv:2408.03335*, 2024.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2019.

[13] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability," *Front. Comput. Sci.*, vol. 7, 2025, doi: 10.3389/fcomp.2025.1520741.

[14] S. H. Mohammed et al., "Dual-hybrid intrusion detection system to detect false data injection in smart grids," *PLoS ONE*, vol. 20, no. 1, Jan. 2025, doi: 10.1371/journal.pone.0316536.

[15] S. Muneer et al., "A critical review of artificial intelligence based approaches in intrusion detection: A comprehensive analysis," *J. Eng.*, vol. 2024, Hindawi, 2024, doi: 10.1155/2024/3909173.

[16] J. Ruan et al., "Deep learning for cybersecurity in smart grids: Review and perspectives," *Energy Convers. Econ.*, vol. 4, no. 4, pp. 233–251, 2023, doi: 10.1049/enc2.12091.

[17] T. Sasilatha, A. A. Suprianto, and H. Hamdani, "AI-driven approaches to power grid management: Achieving efficiency and reliability," *Int. J. Adv. Artif. Intell. Mach. Learn.*, vol. 2, no. 1, pp. 27–37, 2025, doi: 10.58723/ijaaiml.v2i1.380.

[18] F. Ullah, S. Ullah, G. Srivastava, and J. C. W. Lin, "IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic," *Digital Commun. Netw.*, vol. 10, no. 1, pp. 190–204, 2024, doi: 10.1016/j.dcan.2023.03.008.

[19] E. Wagner, L. Bader, K. Wolsing, and M. Serror, "Sherlock: A dataset for process-aware intrusion detection research on power grid networks: Dataset paper," in *Proc. 15th ACM Conf. Data Appl. Secur. Privacy (CODASPY '25)*, 2025, pp. 419–424, doi: 10.1145/3714393.3726006.

[20] T. Yu et al., "An advanced accurate intrusion detection system for smart grid cybersecurity based on evolving machine learning," *Front. Energy Res.*, vol. 10, 2022, doi: 10.3389/fenrg.2022.903370.