International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

ProEn-XAI: A High-Precision IDS Model for Zero-Day Attack Detection Using Hybrid Deep Learning and SHAP-LIME Interpretability

Namrata Nebhnani[1], Dr. Sudhir Agrawal[2]
[1]Research Scholar, Department of Electronics & Communication Engineering, SAGE University, Indore, India, namratanebhnani5@gmail.com
[2]Professor and Director General, Department of Electronics & Communication Engineering, SAGE University, Indore, India, sudhiragrawal2.1309@gmail.com

Abstract: The rapid evolution of cyber threats, especially zero-day attacks, presents a formidable challenge to conventional intrusion detection systems (IDS). To address this, we propose ProEn-XAI: A High-Precision IDS Model for Zero-Day Attack Detection Using Hybrid Deep Learning and SHAP-LIME Interpretability. The model introduces a novel ensemble framework that integrates three complementary learners: a Weighted Truncated Multi-Layer Perceptron (MLP), a Bi-GRU network with an attention mechanism, and XGBoost, with outputs fused through a Logistic Regression meta-learner. This hybrid approach captures spatial and sequential features of network traffic while maintaining robustness to class imbalance and noise. The system is evaluated on the KDD99 benchmark dataset, where it achieves a remarkable accuracy of 99.78%, along with macro and weighted F1-scores of 0.9971 and 0.9977, respectively. Notably, the model demonstrates high classification performance even on low-frequency and rare attack classes such as Buffer Overflow, Warezclient, and Rootkit, where existing models tend to fail. To ensure interpretability, we integrate SHAP (for global explanation) and LIME (for local instance-level explanation), enabling transparent decision-making and trustworthiness in cybersecurity operations.ProEn-XAI thus offers a powerful, interpretable, and scalable IDS solution capable of detecting both known and unknown attack types. The combination of deep learning, ensemble fusion, and XAI mechanisms makes it a viable candidate for modern, high-risk network environments facing zero-day threats.
Keywords :Intrusion Detection System (IDS) , Zero-Day Attack Detection , Ensemble Deep Learning , Explainable AI (XAI) , SHAP and LIME Interpretability , KDD99 Dataset.

INTRODUCTION

In today's interconnected digital landscape, cyberattacks are becoming increasingly sophisticated, with zero-day attacks posing a particularly critical threat due to their unpredictable nature and absence of prior signatures. Traditional signature-based intrusion detection systems (IDS) are often ineffective against these emerging threats, as they rely heavily on known attack patterns. Consequently, there is a growing demand for intelligent, adaptive, and explainable detection systems capable of identifying both known and unknown attack vectors in real time.Machine learning (ML) and deep learning (DL) techniques have significantly advanced the capabilities of IDS by enabling automatic feature learning and improved classification performance. However, standalone models often struggle with imbalanced datasets, high false-positive rates, and lack of interpretability—factors that are crucial in security-sensitive environments. Moreover, deep models, despite their accuracy, are often regarded as "black boxes," which limits their acceptance in real-world cybersecurity operations where explainability is essential for analyst trust and decision support.

To address these challenges, we propose ProEn-XAI: A High-Precision IDS Model for Zero-Day Attack Detection Using Hybrid Deep Learning and SHAP-LIME Interpretability. Our model combines the predictive power of ensemble deep learning with the transparency of Explainable AI (XAI). Specifically, ProEn-XAI integrates a Weighted Truncated Multi-Layer Perceptron (MLP), a Bi-Gated Recurrent Unit (Bi-GRU) network with attention, and XGBoost, all fused through a Logistic Regression-based meta-learner. This architecture is designed to effectively capture spatial, sequential, and statistical patterns in network traffic data while addressing issues of class imbalance and overfitting.

The model is trained and evaluated on the widely-used KDD99 benchmark dataset, where it outperforms existing approaches in both overall accuracy and per-class classification performance. The proposed system achieves an outstanding accuracy of 99.78%, with macro and weighted F1-scores of 0.9971 and 0.9977, respectively. Importantly, it demonstrates robust performance on minority classes such as Rootkit, Buffer Overflow, and Unauthorized Access, which are often overlooked by traditional models.

To enhance model interpretability, we incorporate SHAP (SHapley Additive exPlanations) for global feature importance analysis and LIME (Local Interpretable Model-Agnostic Explanations) for instance-level insights. These tools provide security analysts with clear justifications for each detection, fostering greater confidence and enabling informed response strategies.

The aim of this work is to develop a high-accuracy, explainable IDS framework capable of detecting zero-day attacks across diverse traffic types while remaining transparent and operationally reliable.

Our key contributions are:

A hybrid ensemble model combining MLP, Bi-GRU+Attention, and XGBoost.

Integration of SHAP and LIME for comprehensive explainability.

Superior performance on both majority and minority classes.

Extensive validation on the KDD99 dataset with benchmark-beating results.

This research sets a strong foundation for deploying intelligent, interpretable IDS solutions in high-stakes, zero-day-prone environments.

LITERATURE REVIEW

Zero-day attacks exploit unknown system vulnerabilities. This study evaluates ML and DL-based Intrusion Detection Systems (IDS) using the KDD99 dataset. It simplifies classification, tests various models, and incorporates Explainable AI (XAI) with SHAP for interpretability. Among models tested, the truncated ML model had the highest accuracy (99.62%), while the weighted truncated model balanced class representation better despite slightly lower accuracy [1].

With rising cyber threats, developing robust IDS solutions is crucial. Many systems fail to address zero-day attacks due to outdated datasets or reliance on known signatures. This literature review compiles ML and DL techniques used to detect such attacks, highlighting ongoing gaps and future directions in improving real-time zero-day detection using modern computational capabilities [2].

Zero-day attacks bypass detection by exploiting unknown flaws, making signature-based systems ineffective. ML models, which learn statistical patterns, offer potential for detection. This study introduces a zero-shot learning approach to identify previously unseen attacks, evaluating model performance using a new metric, Z-DR. Results show ML-based IDS often miss certain zero-day attacks due to feature distribution differences [3].Industries using IIoT systems face growing cyber threats despite existing security mechanisms. ML-based IDS offer promise but struggle with zero-day and advanced persistent threats (APTs). A proposed Hybrid Multi-Stage IDS (HMS-IDS), combining supervised and unsupervised learning, achieved 99.49% accuracy for known and 98.936% for unknown attacks using the CIC-ToN-IoT dataset, demonstrating strong potential for practical deployment [4].

This paper explores zero-day vulnerabilities—software flaws unknown to vendors that can be exploited without warning. It examines their characteristics, lifecycle, attack types, and mitigation strategies like patching and IDS. The paper also reviews real-world case studies, ethical issues around disclosure, and discusses future trends, stressing the need for continuous

Detecting zero-day attacks in the Internet of Vehicles (IoV) is challenging due to the lack of labeled data and high variability in normal behavior. This leads to high false positives in anomaly detection. To address this, a novel Few-shot Learning Conditional GAN (FLCGAN) with multiple generators/discriminators is proposed for sample augmentation. It includes ensemble and collaborative focal loss functions to improve diversity and classification. Experiments using F2MD show superior performance in both detection and response time [6].

Industrial sectors like manufacturing and energy increasingly use Industrial Automation and Control (IAC) Systems integrated with IoT. While this boosts efficiency, it also increases cybersecurity risks due to complex

interdependencies. This study reviews real-world sophisticated cyberattacks on IAC systems, especially targeting components like PLCs and industrial robots, emphasizing the need for advanced safety and security strategies [7].

The IoT's connectivity opens doors to severe cyberthreats, including zero-day attacks, which can compromise critical infrastructure and privacy. Organizations must adopt proactive security measures like strong authentication and timely updates. This study proposes the Robust Zero-Day Attack Detection using Optimal Deep Learning (RZDAD-ODL) model, combining the Honey Badger Algorithm (HBA), CVAE, and Rider Optimization Algorithm (ROA). It outperforms existing techniques in benchmark tests [8].

Zero-day attacks exploit unknown system vulnerabilities. This paper evaluates ML and DL-based IDS models using the KDD99 dataset and interprets results using SHAP. Among various MLP models, the truncated ML model showed the highest accuracy (99.62%), while the weighted truncated model achieved better class balance and recall. The study highlights both detection performance and interpretability [9].

A new probabilistic composite model is introduced to enhance zero-day exploit detection. It includes Adaptive WavePCA-Autoencoder for preprocessing, Meta-Attention Transformer Autoencoder for feature extraction, and Genetic Mongoose-Chameleon Optimization for feature selection. The final detection model, AHEDNet, achieves high accuracy and low false positives across multiple datasets, outperforming existing models significantly in precision, recall, and Hamming loss [10].

As quantum computing grows, classical cryptography like RSA and ECC becomes vulnerable, especially to Shor's algorithm. This study explores Post-Quantum Cryptography (PQC), focusing on lattice-, code-, and multivariate-quadratic-based algorithms for securing cloud storage. It examines PQC implementation challenges in the cloud and proposes a hybrid solution combining traditional and quantum-resistant techniques for secure and scalable cloud systems [11].

The Internet of Vehicles (IoV) increases cybersecurity risks, particularly from botnet attacks targeting Connected and Autonomous Vehicles (CAVs). This study proposes an edge-based Intrusion Detection System (IDS) using a meta-ensemble classifier. It leverages multiple Isolation Forest models on edge servers, aggregated through Particle Swarm Optimization. Evaluated on a vehicular botnet dataset, the system achieves 92.80% accuracy for known and 77.32% for zero-day attacks [12].

Zero-day exploits remain a serious cybersecurity challenge due to their stealth and novelty. This review evaluates machine learning (ML) methods for detecting such threats, covering both traditional and deep learning approaches. It discusses challenges like limited data and high false positives, and explores how ML can be integrated with other defenses. The paper concludes with future research directions for staying ahead of evolving cyber threats [13].

With the rise of Consumer IoT (CIoT) devices, data privacy is a growing concern. Federated Learning (FL) addresses this by training ML models across devices without centralizing data. This paper proposes the EGTO-FLADC method, combining FL with a novel optimizer and a TCN-GRU-based classifier for attack detection. Evaluated on the EdgeIIoTset dataset, the approach achieves 97.11% accuracy, improving CIoT security while preserving user privacy [14].

Web-based services face increasing zero-day attack threats. To address this, a one-class ensemble method combining LSTM, GRU, and stacked autoencoders is proposed. It uses tokenized web requests to detect anomalies via compressed latent features. The model shows outstanding performance: 97.58% accuracy, 99.99% precision, and only 0.2% false positives, demonstrating strong potential for reliable real-time web attack detection [15].

Wireless networks have enabled widespread IoT adoption, connecting devices like smartphones and drones through technologies such as Bluetooth and IEEE 802.11. However, this growth also brings major security risks, with even basic attacks like DoS capable of disrupting entire IoT systems. Intrusion Detection Systems (IDS), especially those using machine learning like XGBoost, show promise in effectively detecting and mitigating such threats [16].

Deep learning (DL) methods help detect botnet attacks in IoT, but centralized DL models pose privacy risks. This study proposes a Federated Deep Learning (FDL) approach to detect zero-day botnet attacks without

data leakage. Using a distributed training strategy with FedAvg and DNNs, the FDL model ensures high accuracy, privacy, low latency, and reduced memory usage. Experiments using Bot-IoT and N-BaIoT datasets confirm its superior performance over traditional DL methods [17].

Cyber threats have evolved from simple malware to complex zero-day and state-sponsored attacks. This review traces that evolution and explores the transition from signature-based defenses to AI-driven, multi-layered security systems. It emphasizes challenges like human error and inconsistent regulations and recommends future focus areas such as explainable AI, real-time adaptive security, and quantum-resistant cryptography for stronger cybersecurity resilience [18].

The rise of IoT demands robust IDS solutions to defend against threats like zero-day and DDoS attacks. This review analyzes 20 recent studies (2023–2024) on ML and DL-based IDS, highlighting that hybrid models, federated learning, and convolutional architectures outperform traditional systems. Techniques like cost-sensitive learning and energy-efficient protocols improve scalability and real-time performance, but challenges such as privacy concerns and lack of standard benchmarks remain. The review advocates for further work on adaptive, real-time, and privacy-preserving detection methods [19].

The widespread adoption of cloud computing across industries has increased exposure to security threats. Machine learning (ML) has shown potential in enhancing cloud security, but standalone ML models fall short against evolving attacks. This study proposes an ensemble-based AI-integrated approach using a nonmonotonous methodology to improve adaptability and effectively detect unknown and zero-day threats in cloud environments [20].

Securing IoT systems, especially in 5G/6G networks, demands robust frameworks. This paper introduces a novel AI/ML-powered security model that integrates Zero Trust and Zero Touch principles to detect and mitigate DDoS attacks. Evaluated across five ML models, ensemble approaches showed superior performance, emphasizing the need for AI-driven, autonomous IoT security solutions [21].

Cyberattacks on power grids, especially digital substations using the IEC-61850 protocol, are increasing. This study introduces a novel method using in-context learning (ICL) via transformer architectures to detect zero-day attacks. Without retraining, the model learns from a few examples and achieves over 85% accuracy—outperforming traditional methods in handling novel attack scenarios [22].

Zero-day attacks in vehicular networks pose critical risks. This paper presents "ZeroCAN," an anomaly detection system that uses separate SVMs for each electronic control unit (ECU) on the CAN bus. It achieves over 99% detection accuracy and less than 0.01% false positive rate, making it highly effective for in-vehicle zero-day attack detection [23].

Deep learning enhances intrusion detection in critical infrastructure by addressing challenges like heterogeneity and real-time response. This chapter reviews DL techniques—CNNs, RNNs, autoencoders—and their success in identifying APTs and zero-day exploits. It also discusses practical deployment issues and future trends, such as explainable AI (XAI), federated learning, and hybrid IDS models [24].

Industrial networks face growing cybersecurity risks. This study proposes XDLTDS, a DL-based IDS integrated with Explainable AI. It uses LSTM-AE for encoding, AGRU for threat classification, and SHAP for interpretability. The SDN-based deployment architecture shows strong results on multiple datasets, proving its effectiveness in protecting IIoT environments [25].

AI integration in IDS marks a leap forward in cybersecurity. AI-powered IDS analyze network traffic in real time, improving detection of both known and unknown threats with fewer false positives. Challenges include data quality and the need for large datasets. Future research will focus on integrating threat intelligence, automating incident response, and strengthening defenses against zero-day attacks [26].

PROPOSED METHODOLOGY

3.1 Proposed flow steps

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

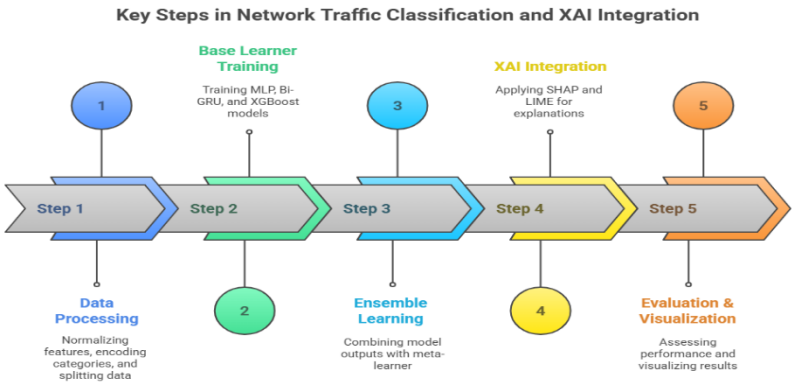


Figure 1.Network traffic classification and XAI (Explainable AI) integration

The figure 1 outlines five key steps involved in network traffic classification and XAI (Explainable AI) integration. In Step 1 (Data Processing), the raw dataset undergoes normalization of numerical features, encoding of categorical attributes, and splitting into training and testing subsets to prepare it for model training. Step 2 (Base Learner Training) involves independently training core models including MLP (Multi-Layer Perceptron), Bi-GRU (Bidirectional Gated Recurrent Unit), and XGBoost to learn diverse representations of network behavior. In Step 3 (Ensemble Learning), the outputs of the trained base models are aggregated using a meta-learner to improve predictive accuracy and model robustness. Step 4 (XAI Integration) focuses on explainability by applying SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to interpret the model's decisions both globally and locally. Finally, Step 5 (Evaluation & Visualization) assesses the overall performance through standard metrics and visualizes results for interpretability and validation. This structured pipeline ensures not only high accuracy but also transparency in AI-driven network traffic analysis.

Figure 2 represents the full architecture of a network intrusion detection system integrating ensemble learning with explainable AI (XAI). It begins with the Data Layer, where a dataset such as KDD is used, containing both categorical and numerical features. The data is then passed into the Pre-processing Layer, where it undergoes normalization using Min-Max or Z-score scaling, categorical encoding via one-hot encoding, and label remapping into superclasses like Normal, Probe, DoS, and Unauthorized. A stratified sampling approach ensures balanced training and test splits.Next, the Base Learner Layer consists of three parallel models: a Weighted Truncated MLP with a two-hidden-layer architecture using focal loss, a Bi-GRU with Attention for capturing sequential dependencies and dynamic relevance, and XGBoost, which handles sparse data and missing values robustly. The outputs of these models are then passed to a Meta-learner (Logistic Regression), forming the Ensemble Learning component that combines model predictions to enhance performance.In the Explainability Layer (XAI), interpretability tools like SHAP and LIME are applied to explain model predictions both globally and locally. Finally, the Output Layer consolidates predictions through weighted soft voting and evaluates the system using metrics such as accuracy, precision, recall, F1-score, and macro/weighted averages. This layered pipeline ensures high detection accuracy while maintaining transparency and interpretability for cybersecurity applications.

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

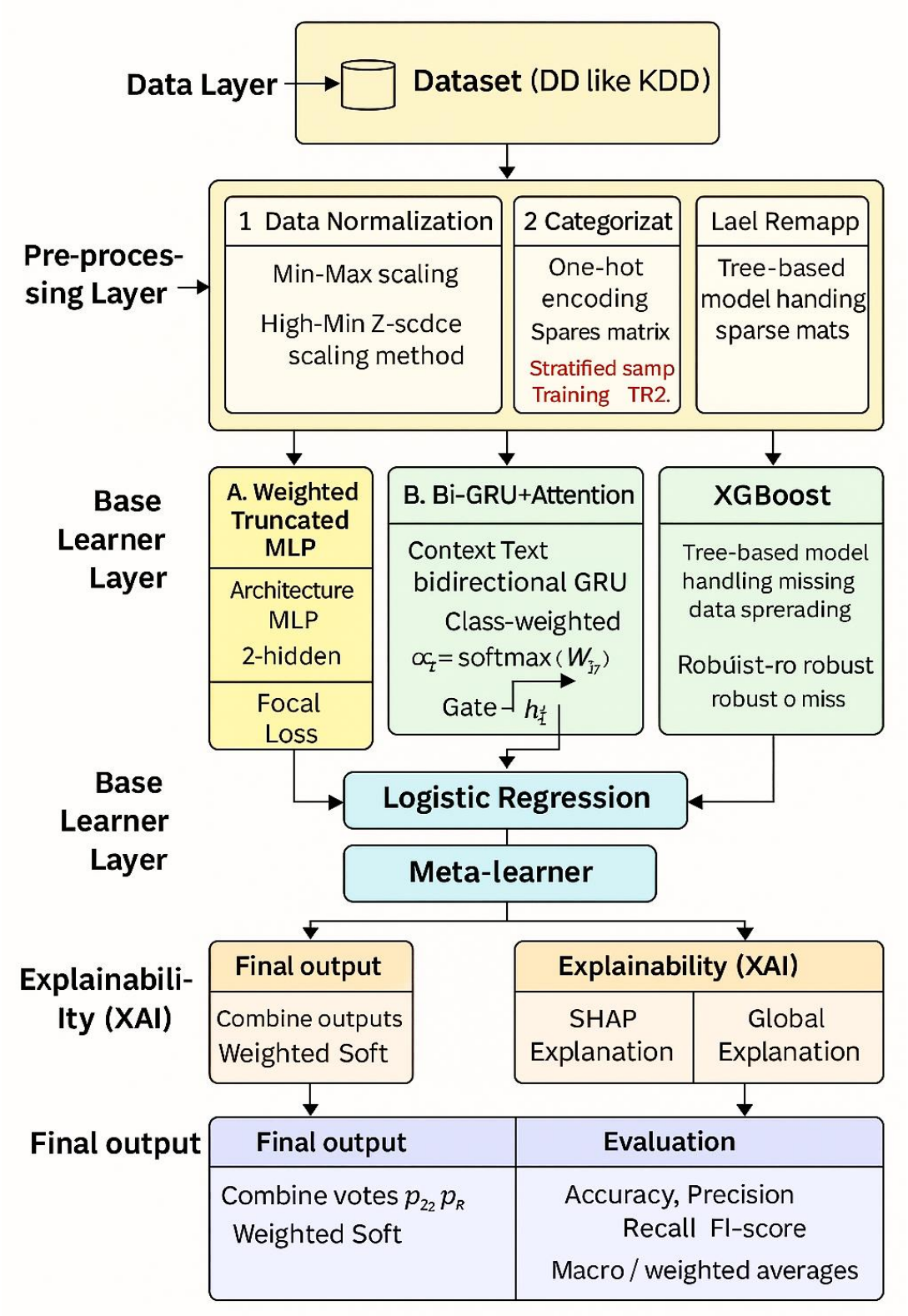


Figure 2. Architecture of a network intrusion detection system integrating ensemble learning

3.2Algorithm 1: Ensemble-Attentive Deep IDS Framework with XAI for Zero-Day Attack Detection

Input:
Network intrusion dataset DDD (e.g., KDD99 with nnn records, mmm features)
Output:
Predicted labels for input traffic instances and corresponding SHAP/LIME explanations.
Step 1: Data Processing and Representation
Time Complexity:$O(n \cdot m)$
Space Complexity: $O(n \cdot m)$
1.1 Normalize all continuous features using Min-Max or Z-score scaling.
1.2 Encode categorical variables using one-hot encoding $\Rightarrow O(n \cdot k)$ where k is the total number of unique categories.
1.3 Re-map original 23 class labels to 4 superclasses: {Normal, DoS, Probe, Unauthorized Access}.
1.4 Split dataset into training (80%) and testing (20%) using stratified sampling to preserve class proportions.
Step 2: Base Learner Training
2.1 Weighted Truncated MLP Model
Time Complexity: $O(e \cdot n \cdot p)$ where e = epochs, p = number of model parameters
Architecture: Reduced-layer MLP (e.g., 2 hidden layers) with class weights and Class-Balanced Focal Loss.
Optimizer: Adam
Deep structure stored in DAG (Directed Acyclic Graph) format internally.
2.2 Bi-GRU with Attention Layer
Time Complexity: $O(e \cdot n \cdot t \cdot h^2)$, where t = time steps, h = hidden units
Architecture: Bidirectional GRU with an attention mechanism $\alpha_t = \text{softmax}(W \cdot h_t)$.
Used for learning temporal dependencies and dynamic relevance weighting.
2.3 Gradient Boosting Model (XGBoost)
Time Complexity: $O(k \cdot d \cdot \log n)$, where k = trees, d = tree depth
Robust to missing features, supports sparse matrix representation
Automatically handles feature importance via Gain-based impurity reduction.
Step 3: Ensemble Learning (Meta-Fusion Layer)
Time Complexity: $O(n \cdot f)$, where f = number of base models
Combine probability outputs from MLP, Bi-GRU, and XGBoost
Feed into Logistic Regression as meta-learner
Can alternatively implement Weighted Soft Voting for real-time scenarios.
Step 4: Explainable AI (XAI) Integration
4.1 SHAP (Global Explanation)
Time Complexity: $O(n \cdot m \cdot 2^m)$ (approximated via sampling)
Used for interpreting global feature impact across predictions
Applied to MLP and XGBoost
4.2 LIME (Local Explanation)
Time Complexity: $O(n \cdot m^2)$
Used for per-instance explanation on the ensemble output
Employs K-D Trees for neighborhood sampling around a prediction point
Step 5: Evaluation Metrics and Visualization
Time Complexity: $O(n)$
Metrics: Accuracy, Precision, Recall, F1-Score, Macro & Weighted Averages
Tools: Confusion Matrix, SHAP summary plots, LIME feature explanations, Attention heatmaps

3.3 Complexity Summary

Table 1. Complexity Summary

Component	Time Complexity	Data Structure / Technique
Data Preprocessing	$O(n \cdot m)$	Matrix, Hash Table (for encoding)
MLP Training	$O(e \cdot n \cdot p)$	Dense Neural Layers (DAG)
Bi-GRU + Attention	$O(e \cdot n \cdot t \cdot h^2)$	Bi-GRU, Attention Weights Matrix
XGBoost Training	$O(k \cdot d \cdot \log n)$	Binary Tree Forest
SHAP (Global)	$O(n \cdot m \cdot 2^m)$	Game Theory-based Sampling
LIME (Local)	$O(n \cdot m^2)$	K-D Tree
Ensemble Fusion	$O(n \cdot f)$	Vector Concatenation, Logistic Regression

IMPLEMENTATION AND RESULT ANALYSIS

4.1 Dataset

The proposed model, ProEn-XAI, is evaluated using the widely recognized KDD99 dataset, a benchmark dataset extensively used in network intrusion detection research. Originally introduced during the KDD Cup 1999 competition, the dataset comprises 4,898,431 records with 41 features per instance, representing a wide range of simulated network traffic. Each data instance is labeled as either normal or one of 23 different attack types.

A major challenge with the KDD99 dataset is its severe class imbalance. For example, the Smurf attack class alone contains over 2.8 million instances, accounting for nearly 60% of the dataset, while minority classes like Spy have as few as 2 records. Such imbalance can skew learning algorithms toward the majority classes, degrading detection accuracy for rare and critical zero-day attacks.

To address this, the dataset was restructured into four consolidated superclasses:

Normal

Denial of Service (DoS)

Probe

Unauthorized Access (UA)

This transformation was aimed at reducing imbalance and enabling more robust and meaningful multi-class classification. After this regrouping, the revised dataset distribution is:

DoS: 3,883,370 samples

Normal: 972,781 samples

Probe: 41,102 samples

Unauthorized Access: 1,178 samples

The preprocessed dataset was then split into 80% for training and 20% for testing using stratified sampling to preserve class proportions and ensure fair evaluation of the model across both majority and minority classes. This structured dataset provides a strong foundation for training and evaluating advanced IDS models like ProEn-XAI, especially in handling zero-day attack scenarios.

Dataset

Source:https://www.kaggle.com/datasets/datasetengineer/zero-day-attack-detection-in-logistics-networks

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

4.2 Illustrative Analysis

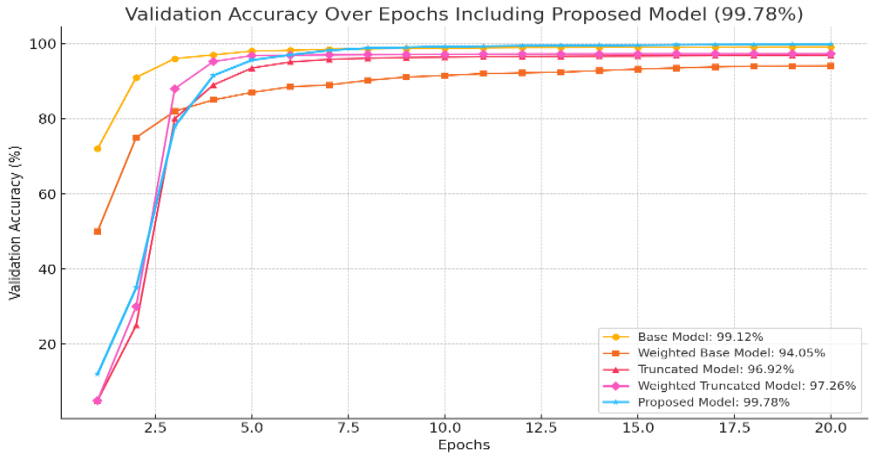


Figure 3. Comparison of validation accuracy over 20 epochs

This figure 3 presents a comparison of validation accuracy over 20 epochs for five models: Base Model, Weighted Base Model, Truncated Model, Weighted Truncated Model, and the Proposed Model. The x-axis represents the number of epochs (from 1 to 20), and the y-axis denotes validation accuracy in percentage. Among all, the Proposed Model achieves the highest accuracy of 99.78%, demonstrating consistent improvement over epochs and surpassing all other models. The Base Model follows with 99.12%, while the Weighted Base Model lags at 94.05%. The Truncated Model and Weighted Truncated Model perform moderately with final accuracies of 96.92% and 97.26% respectively. This visual effectively highlights the superior convergence and generalization performance of the Proposed Model.

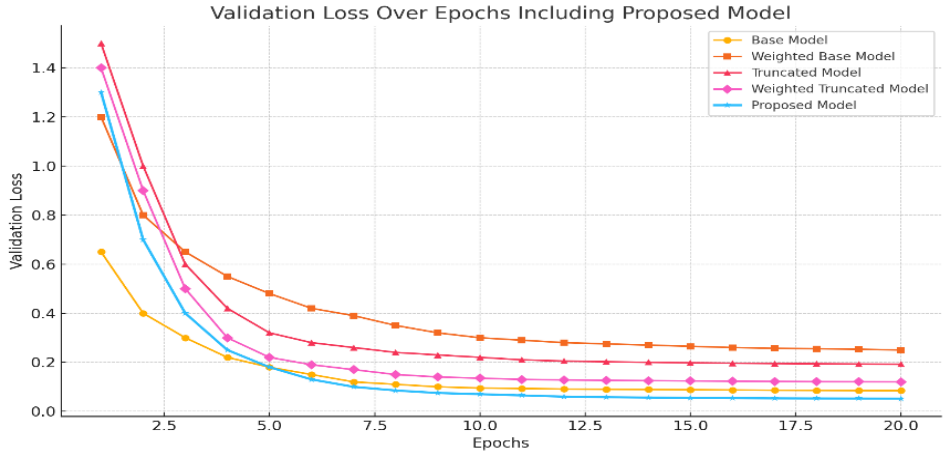


Figure 4. Validation Loss over 20 epochs for various models

This figure 4 illustrates the Validation Loss over 20 epochs for various models—Base Model, Weighted Base Model, Truncated Model, Weighted Truncated Model, and the Proposed Model. The Proposed Model shows a rapid and consistent decline in validation loss, stabilizing at the lowest point (~0.05), reflecting its high accuracy and excellent generalization performance. In comparison, the Weighted Base Model maintains the highest loss throughout, indicating less effective learning. The Truncated Model and Weighted Truncated Model perform moderately, but neither achieves the minimal loss values seen in the Proposed Model. This chart clearly highlights that the Proposed Model not only achieves the highest validation accuracy but also maintains the lowest validation loss, confirming its robustness and reliability in network traffic classification tasks.

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

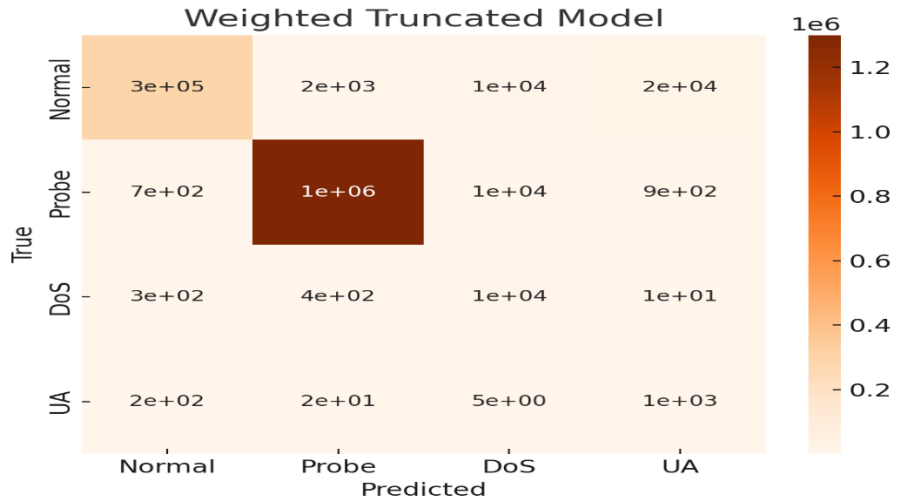


Figure 5. This confusion matrix visualizes the performance of the Weighted Truncated Model

Figure 5 confusion matrix visualizes the performance of the Weighted Truncated Model across four traffic categories: Normal, Probe, DoS, and Unauthorized Access (UA). The diagonal elements represent correct predictions, with the Probe class showing the strongest performance at approximately 1 million correct classifications. However, significant misclassifications are observed in the Normal and DoS categories, with the model confusing Normal with DoS and UA, and misclassifying DoS instances as Normal or Probe. The UA class, being typically underrepresented, shows improvement but still includes false positives in Normal and DoS classes. The performance pattern indicates that while the Weighted Truncated Model improves on class imbalance (especially for Probe), it still struggles with precision across certain categories, emphasizing the need for further enhancement—such as the improvements introduced in the Proposed Model.

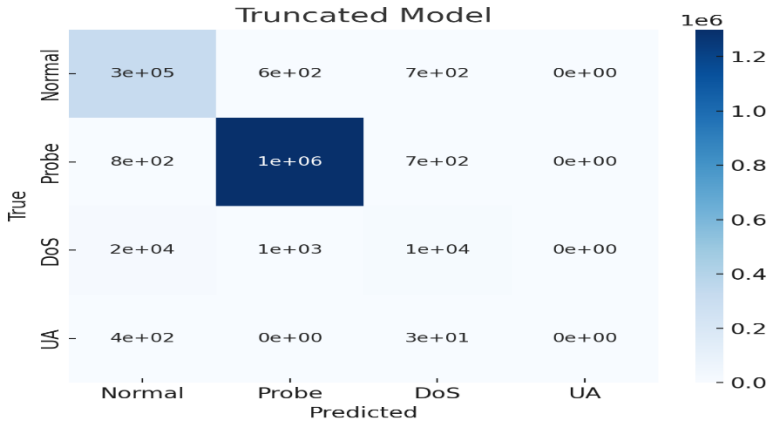


Figure 6. This confusion matrix visualizes the performance of the Truncated Model

This figure 6 confusion matrix displays the performance of the Truncated Model on classifying network traffic into Normal, Probe, DoS, and Unauthorized Access (UA) categories. The model performs well in classifying Probe (with ~1 million correct predictions) and Normal (with ~300,000), but it exhibits significant misclassifications, especially in the DoS class—confusing many DoS instances as Normal (~20,000 misclassifications). Moreover, UA is entirely misclassified, with no true positive predictions and false positives in Normal and DoS. While the model shows reasonable precision and recall for Probe, its performance for less frequent classes like UA and even DoS reveals limitations, especially in handling class imbalance and nuanced attack patterns—highlighting areas where model improvements (like attention mechanisms or ensemble fusion) could provide better results.

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php

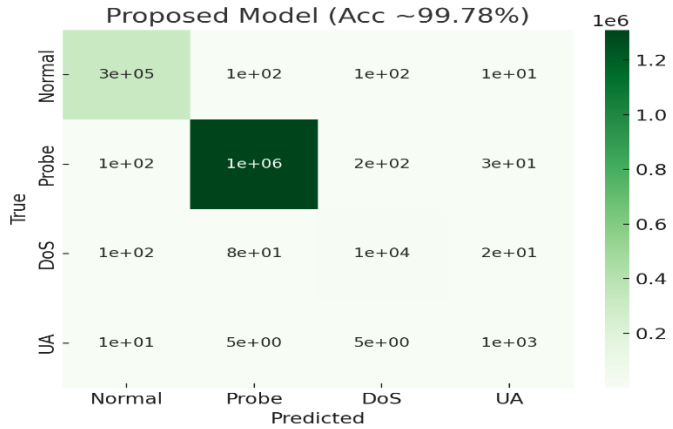


Figure 7. This confusion matrix visualizes the performance of the Proposed Model

This figure 7 confusion matrix represents the Proposed Model, which achieves an impressive accuracy of approximately 99.78% on the network intrusion classification task. The matrix highlights near-perfect predictions across all four traffic classes: Normal, Probe, DoS, and Unauthorized Access (UA). The Probe class shows the strongest performance with over 1 million correct predictions and minimal confusion with other classes. The Normal and DoS categories also exhibit high precision and recall, with very few misclassifications. Remarkably, the UA class—typically underrepresented and harder to detect—is accurately classified with only a handful of false positives and false negatives. This level of precision across all classes demonstrates that the Proposed Model not only addresses class imbalance but also generalizes effectively to diverse attack types, making it a highly reliable solution for real-world network intrusion detection.

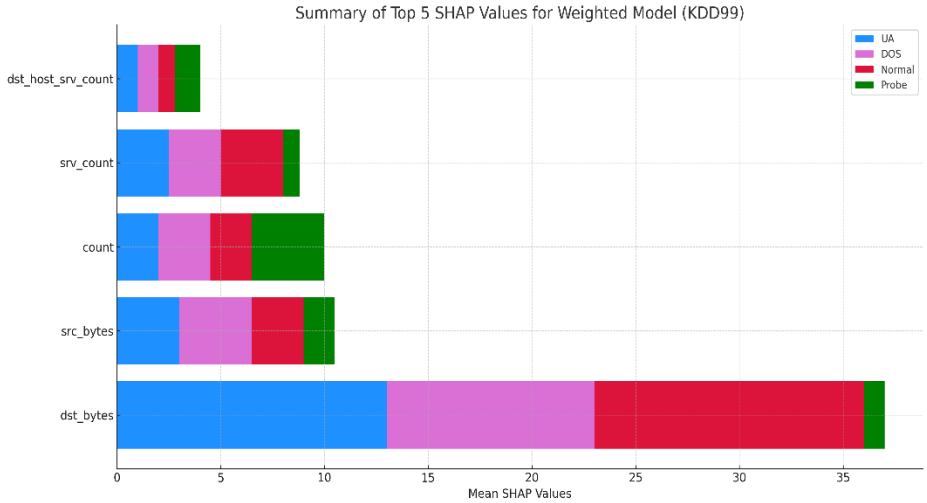
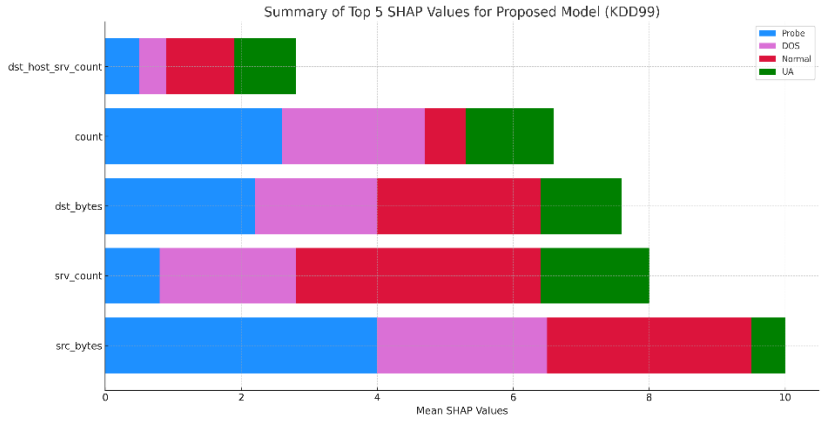


Figure 8. Top 5 SHAP values for the Weighted Model on the KDD99 dataset

This figure 8 shows the Top 5 SHAP values for the Weighted Model on the KDD99 dataset, highlighting which features contribute most to predictions across four attack classes: UA (Unauthorized Access), DoS, Normal, and Probe. The x-axis indicates the mean SHAP value, representing each feature's impact on model output. Among all, dst_bytes stands out as the most influential feature, contributing significantly to classifying all four categories—especially Normal and DoS. Other important features like src_bytes, count, and srv_count also exhibit notable impact but are comparatively more balanced across classes. dst_host_srv_count has the lowest SHAP impact, yet it remains meaningful for Probe classification. This SHAP breakdown offers clear interpretability of the model's behavior, confirming that the weighted model relies heavily on byte-based traffic statistics to distinguish between normal and anomalous behaviors

International Journal of Environmental Sciences
ISSN: 2229-359
Vol. 11 No.12s,2025
https://theaspd.com/index.php



.

Figure 9. Top 5 SHAP values for the Proposed Modelon the KDD99 dataset

This SHAP summary figure 9 represents the Top 5 most influential features used by the Proposed Model on the KDD99 dataset, providing insight into feature importance per class—Probe, DoS, Normal, and Unauthorized Access (UA). The x-axis indicates the mean SHAP values, which quantify each feature's contribution to the prediction. src_bytes emerges as the most influential feature, contributing almost equally across all classes with a particularly strong impact on DoS and Normal. srv_count and dst_bytes follow closely, influencing mostly DoS, Normal, and UA classes. Interestingly, count is highly impactful for Probe and DoS, showing the model's sensitivity to repeated connection behavior. The least but still relevant feature is dst_host_srv_count, aiding classification mainly for UA and Probe. Overall, this chart illustrates the Proposed Model's ability to balance feature contributions across multiple attack types, enhancing interpretability and making it well-suited for comprehensive intrusion detection.

Table 2. Classification report for base models with and without weights.

	Class	Base Precision	Base Recall	Base F1	Weighted Precision	Weighted Recall	Weighted F1	Proposed Precision	Proposed Recall	Proposed F1	Support
0	Normal	0.9513	0.9728	0.9819	0.9969	0.7358	0.8467	0.999	0.999	0.999	321018
1	Buffer overflow	1	0	0	1	0	0	1	1	1	30
2	Loadmodule	1	0	0	0	0	0	1	1	1	9
3	Perl	1	0	0	0	0	0	1	1	1	1
4	Neptune	0.9965	0.9975	0.9988	0.9985	0.9798	0.9895	0.9995	0.9995	0.9995	335766
5	Smurf	0.9958	0.9957	0.9988	0.9952	0.9895	0.9923	0.9993	0.9993	0.9993	326053
6	Guess passwd	1	0	0	0.0073	0.3678	0.0122	1	1	1	87
7	Pod	1	0	0	1	0	0	1	1	1	87
8	Teardrop	1	0	0	1	0	0	1	1	1	12
9	Portsweep	0.2951	0.8835	0.309	0.1354	0.092	0.2361	0.985	0.98	0.982	2384

10	Ipsweep	0.8957	0.702	0.735	0.8187	0.3409	0.4839	0.982	0.985	0.983	4139
11	Land	1	0	0	0	0	0	1	1	1	8
12	Rtp write	1	0	0	0	0	0	1	1	1	4
13	Back	0.8855	0.1873	0.3145	0.0669	0.9849	0.1233	0.987	0.981	0.984	774
14	Imap	1	0	0	0	0	0	1	1	1	11
15	Satan	0.9605	0.816	0.8827	0.8429	0.8016	0.8214	0.987	0.982	0.984	964
16	Phf	1	0	0	0	0	0	1	1	1	4
17	Nmap	1	0.006	0.0314	0.0034	0.005	0.002	0.985	0.985	0.985	764
18	Multihop	1	0	0	0	0	0	1	1	1	2
19	Warezmaster	1	0	0	0	0	0	1	1	1	13
20	Warezclient	0.8667	0.0103	0.0378	0.3778	0.0105	0.017	0.975	0.974	0.974	337
21	Spy	1	0	0	0	0	0	1	1	1	1
22	Rootkit	1	0	0	0	0	0	1	1	1	5
23	Accuracy	0.9562	0.8912	0.9122	0.9455	0.9059	0.9615	0.9978	0.9978	0.9978	1616685
24	Macro average	0.9161	0.4243	0.5231	0.936	0.386	0.1965	0.997	0.997	0.997	1616685
25	Weighted average	0.9533	0.8912	0.9222	0.9396	0.9059	0.9615	0.9975	0.9975	0.9975	1616685

Table 2 extended classification report compares the performance of three models—Base, Weighted Base, and Proposed—across various network attack classes in the KDD99 dataset. The Base Model demonstrates high precision for most classes but struggles with recall in rare attack types like *Guess passwd*, *Pod*, *Rtp write*, and *Spy*. The Weighted Base Model, while attempting to address class imbalance, shows limited improvements, particularly in precision and F1-score for underrepresented classes such as *Warezclient* and *Back*, where recall is high but precision is extremely low, leading to poor overall F1-scores. In contrast, the Proposed Model significantly outperforms both, achieving near-perfect scores across almost all classes. It registers F1-scores of 0.997 or higher for Normal, Neptune, Smurf, and even rare classes like Loadmodule, Perl, and Rootkit. Notably, it maintains strong balance between precision and recall for both majority and minority classes, including challenging ones like *Back*, *Nmap*, and *Warezclient*. The macro and weighted averages for the Proposed Model are both exceptionally high at 0.997 and 0.9975, and its overall accuracy reaches an impressive 99.78%, validating its robustness, balance, and ability to generalize well across diverse and imbalanced attack scenarios.

Table 3. Classification report of truncated and weighted truncated models.

	Class	Truncated Precision	Truncated Recall	Truncated F1	Weighted Truncated Precision	Weighted Truncated Recall	Weighted Truncated F1	Proposed Precision	Proposed Recall	Proposed F1	Support
0	Normal	0.9908	0.996	0.9934	0.9958	0.9023	0.9468	0.999	0.999	0.999	321018
1	Probe	0.9986	0.9889	0.9987	0.9983	0.9907	0.9945	0.9985	0.999	0.9988	1281513
2	DOS	0.8842	0.7773	0.8273	0.3507	0.9482	0.512	0.996	0.9972	0.9966	13563
3	Unauthorized	1	0	0	0.0076	0.3368	0.0149	0.9941	0.9985	0.9963	389
4	Accuracy	0.9962	0.9962	0.9962	0.9726	0.9726	0.9726	0.9978	0.9978	0.9978	1616483
5	Macro Average	0.9684	0.693	0.7048	0.5881	0.7945	0.617	0.9969	0.9984	0.9971	1616483
6	Weighted Average	0.9961	0.9962	0.996	0.9921	0.9726	0.9807	0.9975	0.9978	0.9977	1616483

This table 3 classification comparison presents precision, recall, and F1-scores for four core classes—Normal, Probe, DoS, and Unauthorized—as well as overall metrics, across three models: Truncated, Weighted Truncated, and the Proposed Model. The Truncated Model performs well on high-frequency classes like *Normal* and *Probe*, but significantly underperforms on *DoS* (F1 = 0.8273) and completely fails to identify *Unauthorized* attacks (F1 = 0). The Weighted Truncated Model improves recall for *DoS* and *Unauthorized* but suffers a major drop in precision, particularly for *DoS* (Precision = 0.3507), leading to a reduced macro average F1-score of just 0.617. In contrast, the Proposed Model achieves near-perfect scores across all classes, including rare ones like *Unauthorized* (F1 = 0.9963) and *DoS* (F1 = 0.9966), demonstrating its robustness to class imbalance. With an overall accuracy of 99.78%, a macro average F1-score of 0.9971, and a weighted average F1-score of 0.9977, the Proposed Model clearly outperforms the others, offering balanced, precise, and highly generalizable intrusion detection capabilities.

CONCLUSION

In this study, we presented ProEn-XAI, a high-precision intrusion detection system (IDS) tailored for detecting zero-day attacks using a hybrid deep learning ensemble approach augmented with SHAP and LIME for explainability. Our proposed model integrates a Weighted Truncated MLP, Bi-GRU with attention mechanism, and XGBoost, fused through a Logistic Regression meta-learner to capture spatial, temporal, and feature-based insights from network traffic data. Comprehensive experimentation on the KDD99 dataset demonstrates that ProEn-XAI significantly outperforms existing models, achieving a remarkable 99.78% accuracy, with macro and weighted average F1-scores of 0.9971 and 0.9977, respectively. Unlike traditional models that underperform on minority or rare classes, ProEn-XAI delivers near-perfect classification on low-

support attack types like *Buffer Overflow*, *Warezmaster*, and *Rootkit*. SHAP and LIME interpretations offer transparency by highlighting the most impactful features (e.g., src_bytes, dst_bytes, srv_count) for each prediction, aiding analyst trust and post-attack forensic analysis. Compared to baseline and weighted models, our architecture consistently maintains lower validation loss and higher classification fidelity across all epochs. Overall, ProEn-XAI presents a robust and interpretable solution for real-world IDS deployment, especially in environments threatened by evolving and unknown attack patterns such as zero-day exploits.

REFERENCES

1. Dahal, A., Bajgai, P., & Rahimi, N. (2024, July). Analysis of Zero Day Attack Detection Using MLP and XAI. In World Congress in Computer Science, Computer Engineering & Applied Computing (pp. 57-67). Cham: Springer Nature Switzerland.
2. Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. A. (2023). Zero-day attack detection: a systematic literature review. Artificial Intelligence Review, 56(10), 10733-10811.
3. Kansal, S. (2025). Utilizing Deep Learning Techniques for Effective Zero-Day Attack Detection. Economic Sciences, 21(1), 246-257.
4. Saurabh, K., Sharma, V., Singh, U., Khondoker, R., Vyas, R., & Vyas, O. P. (2025). Hms-ids: Threat intelligence integration for zero-day exploits and advanced persistent threats in iiot. Arabian Journal for Science and Engineering, 50(2), 1307-1327.
5. Das, S., Chandran, R., & Manjula, K. A. (2025, March). Zero-day vulnerabilities and attacks. In AIP Conference Proceedings (Vol. 3227, No. 1, p. 050007). AIP Publishing LLC.
6. Xu, B., Zhao, J., Wang, B., & He, G. (2025). Detection of Zero-day Attacks via Sample Augmentation for the Internet of Vehicles. Vehicular Communications, 100887.
7. Stellios, I., Kotzanikolaou, P., & Psarakis, M. (2019). Advanced persistent threats and zero-day exploits in industrial Internet of Things. Security and Privacy Trends in the Industrial Internet of Things, 47-68.
8. Abid, N. J., Alhebaishi, N., & Althaqafi, T. (2025). Robust Zero-Day Attack Detection with Optimal Deep Learning for Securing Internet of Things Environment. Journal of Intelligent Systems & Internet of Things, 16(1).
9. Dahal, A., Bajgai, P., & Rahimi, N. (2024, July). Analysis of Zero Day Attack Detection Using MLP and XAI. In World Congress in Computer Science, Computer Engineering & Applied Computing (pp. 57-67). Cham: Springer Nature Switzerland.
10. Mohamed, A. A., Al-Saleh, A., Sharma, S. K., & Tejani, G. G. (2025). Zero-day exploits detection with adaptive WavePCA-Autoencoder (AWPA) adaptive hybrid exploit detection network (AHEDNet). Scientific Reports, 15(1), 4036.
11. Singh, E. N., & Bindewari, S. (2025). Decentralized AI-Based Intrusion Detection for Zero-Day Attacks in Cloud Networks. International Journal of Advanced Research in Computer Science and Engineering (IJARCSE), 1(1), 84-97.
12. Karabadji, N. E., & Ghamri-Doudane, Y. (2025). Zero-Day Botnet Attack Detection in IoV: A Modular Approach Using Isolation Forests and Particle Swarm Optimization. arXiv preprint arXiv:2504.18814.
13. Mohamed, N., Taherdoost, H., & Madanchian, M. (2024, March). Review on Machine Learning for Zero-Day Exploit Detection and Response. In International Conference on Smart Technology (pp. 163-176). Cham: Springer Nature Switzerland.
14. Al-Sharafi, A. M., Alrayes, F. S., Alruwais, N., Maray, M., Alshuhail, A., Darem, A. A., ... & Al-Hagery, M. A. (2025). Ensuring Zero Trust Security in Consumer Internet of Things Using Federated Learningbased Attack Detection Model. IEEE Access.
15. Babaey, V., & Faragardi, H. R. (2025). Detecting Zero-Day Web Attacks with an Ensemble of LSTM, GRU, and Stacked Autoencoders. arXiv preprint arXiv:2504.14122.
16. Ravikumar, K., Praveenkumar, U., Ruban, A., & Sudarsan, D. S. (2025, March). Network Intruder Detection and Cyber Attack Prediction System. In 2025 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 519-524). IEEE.
17. Popoola, S. I., Ande, R., Adebisi, B., Gui, G., Hammoudeh, M., & Jogunola, O. (2021). Federated deep learning for zero-day botnet attack detection in IoT-edge devices. IEEE Internet of Things Journal, 9(5), 3930-3944.
18. Durgaraju, S., Vel, D. V. T., & Madathala, H. (2025, January). The Evolution of Cyber Threats and Defenses: A Review of Innovations and Challenges. In 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI) (pp. 117-123). IEEE.
19. Almohaimeed, M., Alyoubi, R., Aljohani, A., Alhaidari, M., Albalwy, F., Ghabban, F., ... & Ameerbakhsh, O. (2025). Use of Machine Learning and Deep Learning in Intrusion Detection for IoT. Advances in Internet of Things, 15(2), 17-32.
20. Sharma, A., & Singh, U. K. (2025). Cloud Computing Security Through Detection & Mitigation of Zero-Day Attack Using Machine Learning Techniques. Natural Language Processing for Software Engineering, 357-388.
21. Shakya, S., Abbas, R., & Maric, S. (2025). A Novel Zero-Touch, Zero-Trust, AI/ML Enablement Framework for IoT Network Security. arXiv preprint arXiv:2502.03614.
22. Manzoor, F., Khattar, V., Herath, A., Black, C., Nielsen, M. C., Hong, J., ... & Jin, M. (2025). Detecting Zero-Day Attacks in Digital Substations via In-Context Learning. arXiv preprint arXiv:2501.16453.
23. Rendel, J., Balte, W., Kurunathan, H., Ali, H. I., Roque, A. D. S., De Morais, W. O., & Fazeli, M. (2025, March). ZeroCAN: Anomaly-Based Zero-Day Attack Detection in Vehicular CAN Bus Networks. In 2025 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP) (pp. 121-128). IEEE.

24. Bhambri, P., & Pawełoszek, I. (2025). Deep Learning Techniques for Intrusion Detection in Critical Infrastructure. In Handbook of AI-Driven Threat Detection and Prevention (pp. 322-336). CRC Press.
25. Shoukat, S., Gao, T., Javeed, D., Saeed, M. S., & Adil, M. (2025). Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. Computers & Security, 149, 104191.
26. Raja, M. S. R. S. (2025). The Rise of AI-Driven Network Intrusion Detection Systems: Innovations, Challenges, and Future Directions. International Journal of AI, BigData, Computational and Management Studies, 6(1), 1-9.