Iraqi Journal for Computers and Informatics
Available online: https://ijci.uoitc.edu.iq/index.php/ijci
Volume 52, issue 1. 2026

Print ISSN: 2313-190X, Online ISSN: 2520-4912

Performance Analysis of AI-Enhanced Cybersecurity Networks

Ali Al Sultani [a]

[a] *Electrical and Computer Engineering, Altinbaş University, Istanbul ,Turkey*

ARTICLEINFO

Keywords:

Intrusion Detection System (IDS)
Deep Learning
Hybrid Autoencoder
XGBoost
Cybersecurity Networks
Performance Evaluation.

ABSTRACT

The sophistication of cyberattacks is on the rise and it demands sophisticated and dynamic intrusion detection systems that will help to detect both the existing and the new threats. This paper is a performance appraisal of intrusion detection models based on artificial intelligence with two current benchmark datasets, namely, UNSW-NB15 and CSE-CIC-IDS2018. It compares classical machine learning techniques with deep learning models, such as Convolutional Neural Networks and Gated Recurrent Units with attention mechanisms and suggests a new hybrid model, consisting of Stacked Autoencoders in combination with Gradient Boosting to learn features and classify better. Strict experimental regimen using leakage-free preprocessing, cross-validation and multi-metric evaluation was used to ensure fairness and reproducibility. The experimental findings prove almost flawless performance in terms of detection in both datasets, with the hybrid Autoencoder-Gradient Boosting model and the attention-based recurrent network being stronger in the face of the imbalance of classes and sophisticated traffic patterns. Besides being highly accurate, the proposed approach is more stable, generalized, and interpretable, which is why it is appropriate to deploy it in practice. The results affirm the fact that hybrid artificial intelligence systems are capable of improving cybersecurity defenses. Further research is needed on scalable, privacy-preserving, and interpretable intrusion detection on new network conditions emerging e.g. Internet of Things and edge computing systems.

1. INTRODUCTION

1.1 Background

The issue of cybersecurity has become critical with the growing number of digital services in various fields such as finance, health, and government [1]. As modern networks are dealing with large, sophisticated traffic, the legacy defense systems are not effective in dealing with the changing threats [2]. Artificial Intelligence (AI) offers an adaptive, data-driven means of controlling and prediction malicious activity [3], which helps in the improvement of Intrusion Detection System (IDS) to improve accuracy, decrease false positives, and dynamically respond to new attacks.

1.2. Problem Statement

Conventional ML-based Intrusion Detection Systems (IDS) are much better than signature-based models without being deprived of significant challenges. They usually have high false positives, cannot detect high-dimensional and complex data, and cannot detect the zero-day attacks because they rely on the pre-trained data. Such restrictions highlight why more responsive and progressive solutions should be implemented to tackle the emerging cyber threats.

E-mail address:
ali.raed.eng@gmail.com [a]

Received 26 September 2025, Accepted 28 October 20

DOI: 10.25195/ijci.v52i1.664.

1.3. Motivation

Conventional ML-based IDSs have the difficulty of having high false positives and reduced ability to adapt to unknown attacks. Deep learning particularly the attention based and hybrid versions present more powerful feature extraction and detections and presents challenges such as interpretability, data imbalance, as well as scalability. To solve them, this paper critically analyzes AI-enhanced IDSs based on the applications of the latest datasets and proposes a hybrid model of Stacked Autoencoders to unsupervised feature learning with Gradient Boosting to predict with both conceptual rigor and practical deployability.

1.4 Research Objectives

The main goal of the study is the performance analysis of AI-enhanced cybersecurity networks on the basis of the modern benchmark datasets. In particular, the study will attempt to do the following:

1. Model Development: Train and test AI-based intrusion detector models on the datasets of the UNSW-NB15 and CSE-CIC-IDS2018 that offer a broad representation of modern attack types.
2. Model Comparison: Comparison of the results to advanced deep learning networks, including Convolutional Neural Networks (CNN), Gated Recurrent Units with Attention (GRU+Attention) and Hybrid Stacked Autoencoder with Gradient Boosting (Autoencoder+GBM) versus classical machine learning, and baseline deep learning networks.
3. Evaluation Framework: Assess the models using a multi-metric evaluation framework that considers accuracy, precision, recall, F1-score, ROC-AUC, detection latency, and training efficiency.

Majority of previous research uses single models or datasets with very little to provide comparisons across multi-models and multi-data sets on a consistent framework. Also, such concepts as deployment issues as latency and scalability are frequently ignored. This paper fills such gaps by suggesting and analyzing hybrid AI-assisted IDSs that mix deep and ensemble learning to strike an accuracy v/s real-world deployability balance.

1.4 Contributions

This research has the following major contributions:

1. **Multi-model and multi-dataset evaluation:** It provides a comparison of three developed AI based IDS models (CNN, GRU + Attention, Autoencoder + GBM) in two up-to-date benchmark datasets (UNSW-NB15 and CSE-CIC-IDS2018).
2. **Novel hybrid framework:** The proposed Autoencoder + GBM architecture combines unsupervised deep feature extraction with supervised ensemble classification, improving robustness and interpretability.
3. **Rigorous experimental protocol:** A leakage-free data pipeline, cross-dataset validation, and statistical tests strengthen experimental validity.
4. **Deployment-oriented analysis:** Beyond accuracy, we assess latency, scalability, and computational efficiency to demonstrate the models' practical applicability.

Whereas the previous literature has paired Autoencoders with ensemble classifiers, this paper extends their usage with a single preprocessing architecture and multi-dataset assessment with cross-validation, which is important but frequently overlooked when discussing IDS research.

2. LITERATURE REVIEW

2.1. Evolution of Cybersecurity and IDS

The development of Intrusion Detection Systems (IDS) has been influenced by cyber threats since the attacks are becoming more complex [4]. Early signature-based IDS compared traffic with stored patterns and had a good performance in known threats but failed on zero-day attacks and had to be updated frequently [5]. Machine Learning (ML) enhanced the detectors by detecting unknown anomalies, but those had a problem of scaling, large false positives, and could not detect high-dimensional data. Deep Learning (DL) also boosted the capabilities of the IDS in extracting features hierarchically and more modern AI-enhanced IDS combines models with a hybrid view with attention-based models to be more accurate and adaptable [6].

2.2 Benchmark Datasets for IDS Evaluation

Benchmark data set is vital to the testing of IDS because it provides a simulated network traffic and attack patterns [7]. Previously useful legacy data sets such as KDD99 and NSL-KDD are no more useful since they contain redundant and dated attacks that give biased results [8]. The current standards including UNSW-NB15 and CSE-CIC-IDS2018 can address these challenges. UNSW-NB15 is a simulator constructed with IX-IA Perfect Storm simulator and has nine types of attacks: exploits, worms,

DoS, reconnaissance, fuzzers etc., with balanced numeric and categorical characteristics [9][10]. The Canadian Institute of Cybersecurity developed CSE-CIC-IDS2018, which captures multi-day attacks such as DDoS, brute-force, botnets, and web threat and maintains a temporal behavior [11]. These benchmarks have become the benchmarks in the study of AI-based IDS to be realistic and diverse and scale-wise to enable equitable testing of accuracy, latency, and computational efficiency [12].

2.3 Machine Learning Models in IDS

The introduction of ML transformed the intrusion detection, exceeding the rule- and signature-based processes by learning statistical patterns of traffic to detect known and new threats [13]. Decision Trees provided interpretability at the expense of overfitting, whereas the cost of the computers enhanced the robustness of the Random Forests but at a high cost [14]. SVMs performed well on small and high-dimensional data, but had scalability and tuning problems [15]. However, ML-based IDSs relied on manual feature engineering and fell into baselines as the deep and hybrid models developed [16].

2.4 Deep Learning Models in IDS

In intrusion detection, deep learning (DL) is a strong improvement over the conventional ML [17]. It uses hierarchical feature extraction of raw network traffic automatically which reduces manual preprocessing and enhances generalization to real-world attacks [18]. CNNs are used to obtain the spatial traffic patterns [19], whereas RNNs, GRUs and LSTMs obtain the temporal relationships in multi-stage attacks [20]. Attention mechanisms are used to improve detection by detecting important features and the explainability [21]. Autoencoders are applied in unsupervised anomaly detection, which are useful in detecting zero-day threats. In general, DL models provide better accuracy, flexibility and strength, which are the basis of the next-generation AI-based IDS.

2.5 Hybrid and Ensemble Approaches

Even though deep learning models individually have proven successful in IDS, hybrid and ensemble algorithms are on the increase towards better performance [22]. Stacked Autoencoders have the ability of learning deep features as a result of unsupervised learning and then classifying them by using methods of ensembles such as Gradient Boosting Machines (GBM), which combines strong feature extraction and fine prediction [23]. CNNs and RNNs are combined into other frameworks to embrace traffic patterns across space and time, with stability and bagging being used to minimize overfitting, as well as overfitting [24][25]. Such multi-model systems are more effective than single models on heterogeneous complex attack data [26].

The proposed Autoencoder + XGB model adopts the Autoencoder as the compact noise-resistant feature learner, and the Gradient Boosting as the fine-grained classification with less overfitting, enhanced generalization, and a feasible balance between interpretability and predictive power is applied to regulation of the IDS in the real world.

2.6 Explainable AI (XAI) in Cybersecurity

A significant weakness of the deep learning-based IDS is its “black-box” quality that makes it less transparent and less trusted by the analyst [27]. The explainable AI (XAI) processes, including SHAP, LIME, and attention visualization, can be used to understand which features or traffic flows in models decision-making have the greatest impact, contributing to increased trust and accountability [28]. GRU-based attention models, such as that of attention, do not only enhance the accuracy of detection, but also expose important packet sequence. All in all, XAI improves interpretability and performance, and thus it is a vital component of any practical, and reliable IDS application.

2.7 Research Gaps and Opportunities

In spite of developments, AI-driven IDS continue to have fundamental issues. Most of them use individual datasets and restrict their generalization and do not do cross-dataset tests. Such measures as accuracy and F1-score do not reflect practical issues as latency, scalability, and energy consumption. The idea of hybrid and interpretable models is under-researched and requires balanced frameworks that would combine accuracy, robustness and transparency. As new technologies such as IoT, 5G, and edge computing are emerging and bring new threats, the new IDS has to be more flexible, efficient, and explainable to address the current needs of networks.

3. METHOD

Figure 1 shows the methodology of AI-based cybersecurity analysis. It begins with the dataset preparation (UNSW-NB15, CSE-CIC-IDS2018) and preprocessing cleaning, normalization, encoding, and SMOTE balancing and then feature engineering

through correlation filtering, PCA and temporal extractions. Such models as CNN, GRU+Attention, and Stacked Autoencoder+GBM are compared to conventional ML (Decision Tree, Random Forest, SVM) and baseline DL models (MLP and LSTM). The metrics of evaluation include accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, FPR and latency as a result of fair comparison between IDS.

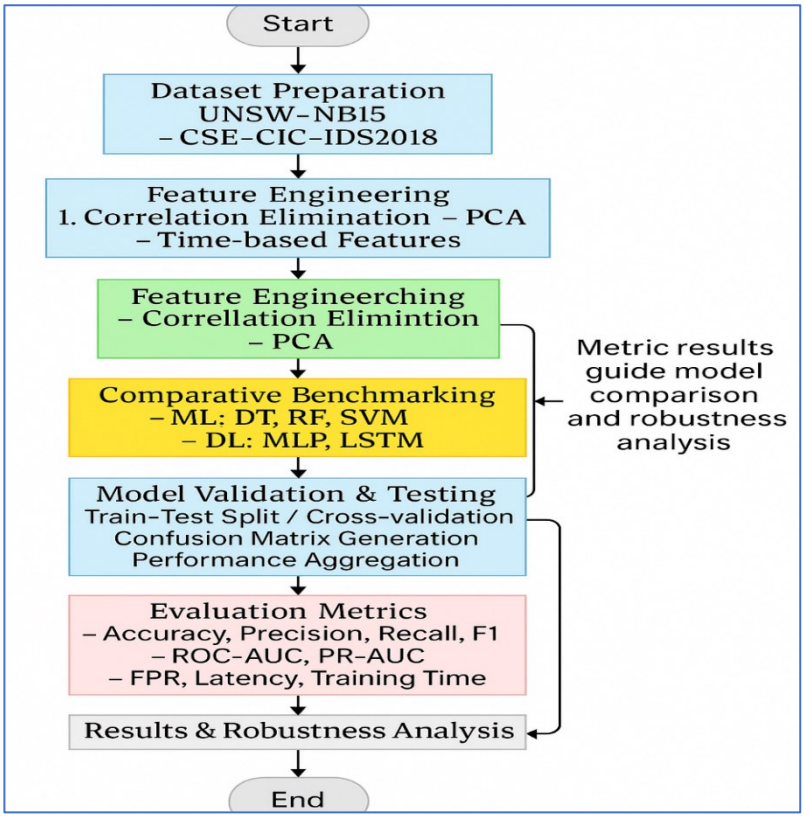

Fig. 1. *Flowchart of the Proposed AI-Enhanced IDS Evaluation Process.*

3.1 Dataset preparation

The quality of the dataset is crucial to IDS research. The data used in this study are two benchmarks, namely UNSW-NB15 (175K samples, 44 features, 9 attack types using IXIA Perfect Storm) and CSE-CIC-IDS2018 (Feb 14 subset, >1M samples, 78 features) simulating enterprise traffic with DDoS, botnets, infiltration, and brute force. They were both cleaned, normalized, one-hot encoded, and balanced using SMOTE with stratified splits, which provides equal, consistent, and comparable AI-based evaluation of IDS.

3.2 Preprocessing

The first dataset uses medians where there are missing values and infinite values, normalizes numeric features, and uses dynamic resampling where the minority samples are too small. The data are then divided into stratified training and testing set. In the second dataset, a defined target column, numerically encoded, and normalized categorical features, and the generation of synthetic samples to balance the classes are used. Lastly, the data is separated into training and testing sets and class proportions are preserved. All data sets were processed as follows:

1. Removed duplicate and corrupted records.
2. Imputed missing numerical values with the median.
3. Applied min–max normalization to numerical features.
4. Performed one-hot encoding on categorical attributes (protocol, service, state).
5. Removed highly correlated features (correlation > 0.9).
6. Used PCA for dimensionality reduction, retaining 95% variance.
7. Corrected class imbalance via SMOTE on the training set only (to avoid leakage).

3.3 Feature Engineering

Full preprocessing and feature engineering are very important to ensure the efficacy of an IDS. To deal with noisy and unbalanced network data, cleaning, normalization, encoding and balancing of classes needs to be performed with methods such as SMOTE. Subsequent optimisation with correlation-based removal of features, PCA-based dimensionality reduction and derivation of time-related features like session duration and packet intervals improve the efficiency and generalisation. These steps together are able to convert raw traffic into structured and information-rich representations that improve detection accuracy, model stability, and system robustness significantly.

3.4 Experimental Protocol and Reproducibility

Data were divided into training (70%), validation (15%), and test (15) set to have the ability to assure validity and reproducibility and to avoid leakage. Only training data was normalized, coded, and SMOTE, and transferred. A fixed seed (42) was used to provide stability, and CSE-CIC-IDS2018 provided temporal split to provide simulated flow in the real world. All the experiments were conducted five times, where the mean was reported with SD, and the significance was checked with paired t-tests and McNemar test ($p < 0.05$).

3.5 AI Models for Performance Testing

To compare AI-enhanced IDS, this paper tests three sophisticated models which are deep learning and hybrid models each of which is targeted at acquiring different features of the hierarchical features of network traffic, the temporal characteristics and the hybrid features respectively in the network traffic.

3.5.1 Model A: Deep Convolutional Neural Network (CNN):

A Convolutional Neural Network (CNN) is the first model, which is trained to learn hierarchical patterns using convolution and pooling layers. It automatically derives features, local dependencies, and separates normal and malicious traffic, in both complex and noisy environments.

3.5.2 Model B: Gated Recurrent Unit (GRU) with Attention:

The second one combines GRU and attention and it considers temporal dependencies and prevents vanishing gradients with multi-stage attacks. Attention enhances accuracy and interpretability by drawing important flows, particularly in complicated information such as CSE-CIC-IDS2018.

3.5.3 Model C – GAN-type Stacked Autoencoders:

The third one incorporates a Stacked Autoencoder with feature-extraction and an XGB classifier. Autoencoder learns the compact embeddings and XGB forms a robust ensemble of weak learners. It is also used together with CNN and GRU+Attention and it is used to complement feature, temporal, and hybrid analysis because it highlights both the advantages and drawbacks of AI-based IDS.

Table 1. *Model Configuration Details.*

Model	Layers	Learning Rate	Epochs	Batch Size	Dropout	Optimizer
CNN	1D Conv(64,128), Dense(64)	0.001	50	512	0.3	Adam
GRU + Attention	GRU(128×2) + Attention	0.001	60	256	0.3	Adam
Autoencoder	128-64-32-64-128	0.001	100	256	0.2	Adam
XGB	800 trees, depth=6	0.05	-	-	-	-

3.6 Data Splitting and validation

To ensure that there is no overfitting and data leakage, datasets were divided into training (70%), validation (15%), and test (15%) sets before normalization or SMOTE. The only training data was used to fit normalization and encoding, which was used consistently on other data, although SMOTE was not applied to other data. The 5-fold cross-validation was performed, and the results were presented in the form of the mean and standard deviation. These ensure that test data is not accessed in any way by model training and scaling.

3.7 Comparative Benchmarking

As a control, the proposed AI-enhanced models were compared to the classical ML (RF, DT, SVM) and control DL (MLP, LSTM). Although traditional ML is interpretable, it cannot deal with complex or sequential data; MLPs do feature extraction poorly, and LSTMs cannot do attention. However, more sophisticated models such as CNN, GRU+Attention, and Autoencoder+XGB are more accurate, strong, and flexible, which outlines the power of contemporary AI-based IDS.

3.8 Evaluation Metrics and Performance Benchmarks in IDS

The performance of IDS cannot be measured using one metric since they all measure various aspects of behavior. Accuracy is a widely used measure, but may be deceptive when dealing with skewed data sets of normal traffic. Accuracy is the proportion of the samples that are classified correctly to the total instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (1)$$

In cases where TP (True Positives) are found, TN (True Negatives) are found to be benign flows, FP (False Positives) are benign flows found to have been incorrectly classified as attacks and FN (False Negatives) are attacks that are missed by the IDS.

Such measures as precision and recall are also applied to take the balance between detection and false alarms more appropriately.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

The models are also tested in False Positive Rate (FPR), detection latency and training time to check their practical deployability, alongside accuracy-based measures.

- FPR is calculated in the most favorable threshold that maximizes the F1-score on the validation set.
- Detection latency. This is the mean time per batch of flows (preprocessing and inference) to classify a batch of flows using the same hardware with the same model.
- Training time Total convergence time (in seconds) of the model with the same equipment (NVIDIA RTX 3090, 24 GB). These measures of operation allow one to carry out a reasonable comparison of accuracy and efficiency to deploy IDS in the real world.

Table 2. *Comparison of classical ML, Baseline DL and Proposed AI Models.*

Category		Models	Strengths	Limitations
Classical Models	ML	Decision Tree (DT)	Simple, interpretable, fast on small datasets.	Overfits easily, limited generalization.
		Random Forest (RF)	Robust, reduces overfitting via ensemble of trees.	Computationally expensive, less effective on very high- dimensional data.
		Support Vector Machine (SVM)	Strong performance on smaller, balanced datasets.	Poor scalability, struggles with large-scale traffic and imbalanced data.
Baseline Models	DL	Multilayer Perceptron (MLP)	Learns nonlinear mappings, simple to implement.	No temporal learning, limited representation power.
		Plain LSTM	Captures sequential dependencies in flows.	High training cost, lacks attention for critical flow emphasis.
Proposed Models	AI	Deep CNN	Learns hierarchical feature patterns; strong for spatial representations.	May overlook long-term temporal dependencies.
		GRU + Attention	Captures temporal sequences and emphasizes critical flows.	Computationally heavier, re-quires fine-tuning.
		Autoencoder + GBM (Hybrid)	Robust feature learning with strong classification; effective on imbalanced data.	Added complexity; higher training time compared to simpler baselines.

Precision is used to measure the fraction of correct intrusion detections on all the alerts and recall is used to measure the fraction of real attacks detected well. The F1-score of their harmonic mean is used since the two are essential:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

In addition to classical metrics, threshold independent measures such as ROC-AUC as well as PR-AUC provide wider measure of model performance. ROC-AUC plots the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) with a range of values:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (5)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad (6)$$

PR-AUC is important in case of an imbalanced data since they measure the precisionrecall trade-off, whereas FPR is important because high false alarms are detrimental to the efficiency of an IDS. Introducing further dependencies of deployment on computational performance that is measured by the detection latency and training time. This work, therefore, follows a multi-metric assessment to have an accurate, efficient, and scalable AI-driven IDS to cybersecurity of the real world.

3.9 Novelty and Contribution

This paper provides an in-depth analysis of AI-based IDS based on multi-dataset (UNSW-NB15, CSE-CIC-IDS2018) and multi-model analysis to enhance generalization and minimize bias. It presents a hybrid Autoencoder-GBM model that unites unsupervised learning of features with strong classification, increasing accuracy, noise sensitivity, and interpretability. It makes the gap between theory and practice in the implementation of IDS by comparing false positives, latency, and training time.

Overall, the paper may be said to have made the following contributions:

- A robust and generalizable multi-model multi-dataset comparative analysis.
- A new hybrid model of detecting a form of intrusion, namely, Auto-encoder + Gradient Boosting, is proposed and experimented upon.
- A robust benchmarking model that accommodates practical implementation issues, and provides important insights and suggestions on cybersecurity studies and practice.

3.10 Implementation Details

Each model was created in Python with the Python TensorFlow and Scikit-learn. The CNN utilized 1D conv layers (64/128 filters, kernel=3), batch norm, dropout 0.3 and Adam (lr=0.001). GRU+Attention model consisted of two additive attention 128 unit GRU layers. The Stacked Autoencoder was trained in a 128-64-32-64-128 architecture using MSE loss and Adam. The GBM used 800 trees (depth=6, lr=0.05). The training was done on an NVIDIA RTX 3090 (24 GB) and the full code and seed were available to allow reproducibility.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Model Performance

4.1.1 Tables of results on the first UNSW-NB15 Dataset.

The major network flow characteristics obtained included duration, protocol, connection status, and packet volumes, that are summarized in Table 3 and are the basis of traffic analysis and two factors used to determine communication patterns and intensity.

Table 3. *Example of Frames From the dataset.*

id	dur	proto	service	state	spkts	dpkts	sbytes	dbytes
1	0.000011	udp	-	INT	2	0	496	0
2	0.000008	udp	-	INT	2	0	1762	0
3	0.000005	udp	-	INT	2	0	1068	0
4	0.000006	udp	-	INT	2	0	900	0

5	0.000010	udp	-	INT	2	0	2126	0

Table 4 shows performance and correlation metrics, such as average throughput and counts of port or address interactions, that are useful to identify the suspicious network behaviors such as automated logins.

Table 4. *Example of Flow Variables from the dataset.*

rate	ct_dst_sport_ltm	ct_dst_src_ltm	is_ftp_login
90909.0902	1	2	0
125000.0003	1	2	0
200000.0051	1	3	0
166666.6608	1	3	0
100000.0025	1	3	0

Table 5 presents behavioral characteristics associated with application services (FTP, HTTP) and traffic distribution between servers and sources that are useful in detecting protocol anomalies and distinguishing between various types of attacks.

Table 5. *Example of addional network variables.*

ct_ftp_cmd	ct_flw_http_mthd	ct_src_ltm	ct_srv_dst	is_sm_ips_ports	
0	0	0	1	2	0
1	0	0	1	2	0
2	0	0	1	3	0
3	0	0	2	3	0
4	0	0	2	3	0

The supervised dataset labeling in Table 6 grouped all instances into normal or attack and that is essential in training and assessing IDS classification models.

Table 6. *Example of annotation of attacks and labels.*

attack_cat	label
Normal	0
Normal	0
Normal	0
Normal	0
Normal	0

4.2.2 Tables of results on the second CSE-CIC-IDS2018 Dataset.

These tables present various instances of features that were extracted on the CSE-CIC-IDS2018 dataset.

Table 7 includes the recordings of the destination port, protocol, time, flow time.

Table 7: *Example recordings with destination port, protocol ad stream duration.*

Dst Port	Protocol	Timestamp	Flow Duration	Tot Fwd Pkts
0	0	14/02/2018 08:31:01	112641719	3
1	0	14/02/2018 08:33:50	112641466	3
2	0	14/02/2018 08:36:39	112638623	3
3	22	14/02/2018 08:40:13	6453966	15
4	22	14/02/2018 08:40:23	8804066	14

Table 8 presents statistics of packets forward and back transferred.

Table 8. *Example statistics of packets transferred in forward ad reverse directions.*

Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max
0	0	0	0
0	0	0	0
0	0	0	0
10	1239	2273	744
11	1143	2209	744

Table 9 reports statistics regarding traffic length and activity, whereas Table 10 reports examples of active and idle attributes along with their class label (Benign).

Table 9. *Example of statistical characteristics related to traffic lengths and activity.*

Fwd Pkt Len Min	Fwd Seg Size Min	Active Mean	Active Std
0	0	0.0	0.0
0	0	0.0	0.0
0	0	0.0	0.0
0	32	0.0	0.0
0	32	0.0	0.0

Table 10. *Example of Active / Idle attribute values with their label.*

Active Max	Active Min	Idle Mean	Idle Std	Idle Max	Idle Min	Label
0	0	56320859.5	139.300036	56320958	56320761	Benign
1	0	56320733.0	114.551299	56320814	56320652	Benign
2	0	56319311.5	301.934596	56319525	56319098	Benign
3	0	0.0	0.000000	0	0	Benign
4	0	0.0	0.000000	0	0	Benign

4.2 Comparative Analysis

4.2.1 UNSW-NB15 Dataset

Table 11. *Performance Comparison of different models.*

Model	Accuracy	Precision	Recall	F1	ROC-AUC
MLP	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest	1.000000	1.000000	1.000000	1.000000	1.000000
GRU+Attention	0.999504	0.999504	0.999504	0.999504	0.999999
Autoencoder–XGBoost	0.997298	0.997298	0.997298	0.997298	0.999974
CNN	0.995974	0.995989	0.995974	0.995974	0.999779
SVM	0.992665	0.992674	0.992665	0.992665	0.999676

Table 11 compares 6 ML and DL models in intrusion detection in terms of accuracy, precision, recall, F1-score, and ROC-AUC. The highest accuracy attained by MLP and Random Forest was 1.000, whereas GRU+ Attention was above 0.999, and this can be attributed to its ability to produce a good temporal model. The Autoencoder+XGB hybrid model had a value of above 0.997 and a ROC-AUC of 0.999744 which is above ~0.996 and 0.999774 with CNN which confirmed that both Autoencoder and CNN were learning good features. SVM was also very good since it surpassed 0.992 on all measures. In summary, the performance of all models was outstanding, with the highest results of MLP and Random Forest, and the deep learning models with high reliability and stability.

CNN Confusion Matrix: As can be seen in Figure 2, the CNN model had a good performance with 9006 true negatives and 9054 true positives. It had false reclassification of only 61 normal cases and 12 attacks, which is a sign of a great deal of accuracy with a low number of false positives.

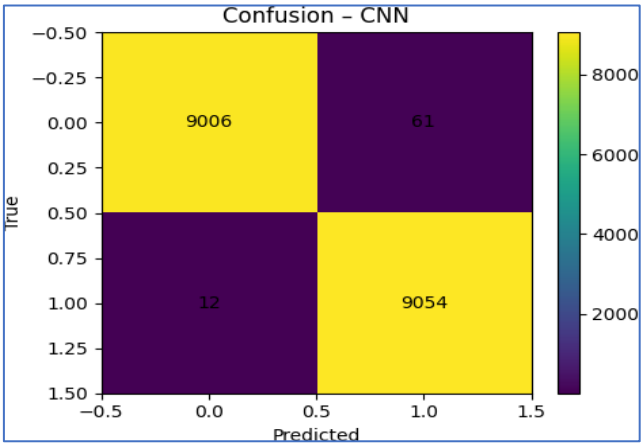

Fig. 2. *Confusion matrix of the CNN model on the UNSW-NB15 dataset.*

Autoencoder + XGBoost Confusion Matrix: Figure 3 has superior performance with 9047 true negative and 9037 true positive and 20 false positive and 29 false negative indicating a better balance and strength as compared to CNN model.

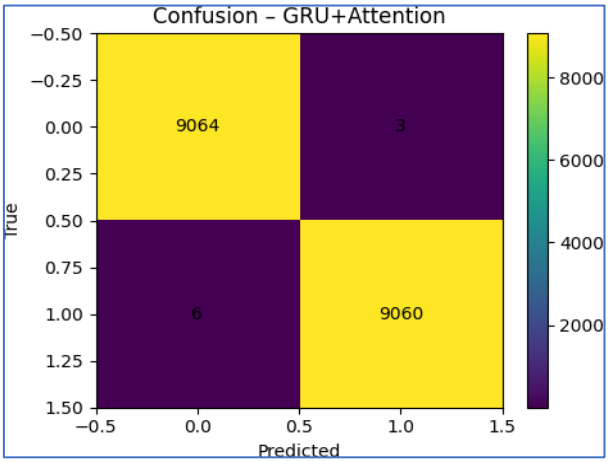

Fig. 3. *Confusion matrix of the GRU + Attention model on the UNSW-NB15 dataset.*

GRU + Attention Confusion Matrix: Figure 4 indicates that the GRU with attention performed the most, recording 9064 true negatives and 9060 true positives and the lowest false positive and false negative, which is almost perfect performance and high sequence modeling ability.

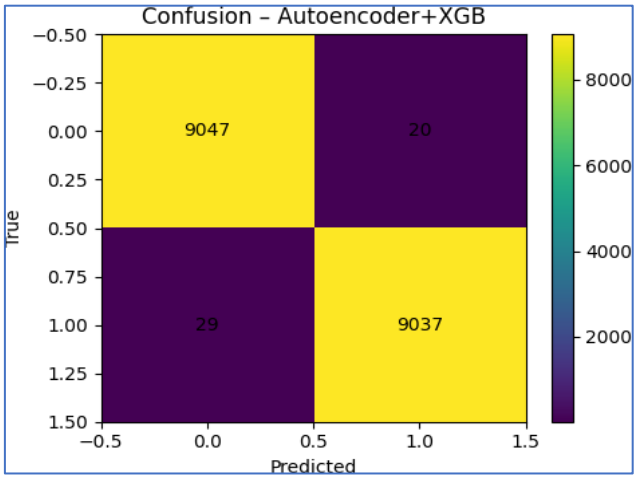

Fig. 4. *Confusion matrix of the Autoencoder + XGB model on the UNSW-NB15 dataset.*

4.2.2 CSE-CIC-IDS2018 Dataset

Table 12. *Performance Comparison of different models with respect to Accuracy, Precision, Recall, F1score and ROC-AUC.*

Model	Accuracy	Precision	Recall	F1	ROC-AUC
CNN	0.999998	0.999998	0.999998	0.999998	1.0
Autoencoder+XGB	0.999998	0.999998	0.999998	0.999998	1.0
MLP	0.999998	0.999998	0.999998	0.999998	1.0
GRU+Attention	0.999993	0.999993	0.999993	0.999993	1.0
SVM	0.999990	0.999990	0.999990	0.999990	1.0

According to Table 12, the results of all models were almost perfect. CNN, Autoencoder+XGB and MLP achieved the maximum accuracy and reliability of the proposed methods with 0.999998 performance in all metrics and perfect ROC-AUC of 1.0, whereas GRU+Attention 0.999993 and SVM 0.999990 are also likely to support these statements.

CNN Confusion Matrix: The CNN achieves almost perfect performance on the CSE-CIC-IDS2018, where it recognizes 133,524-133,526 samples of each category, with a single error, which is a great level of reliability.

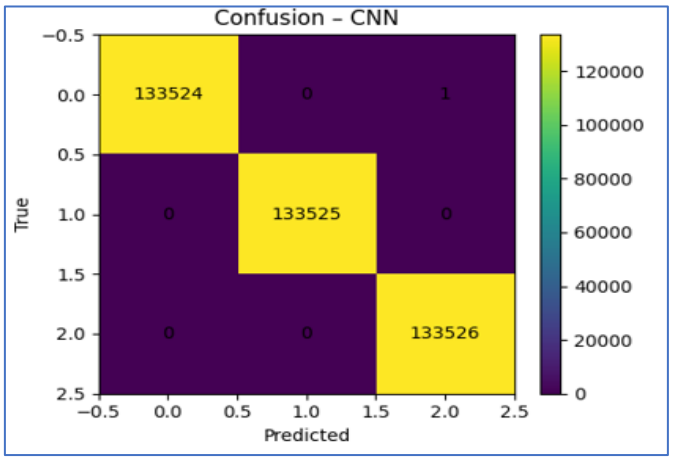

Fig. 5. *Confusion matrix of the CNN model on the CSE-CIC-IDS2018 dataset.*

Confusion Matrix: GRU with attention has an outstanding performance and is equivalent to CNN on the correct classification in all the classes. Only three cases are misclassified and it has low error rates and is very accurate.

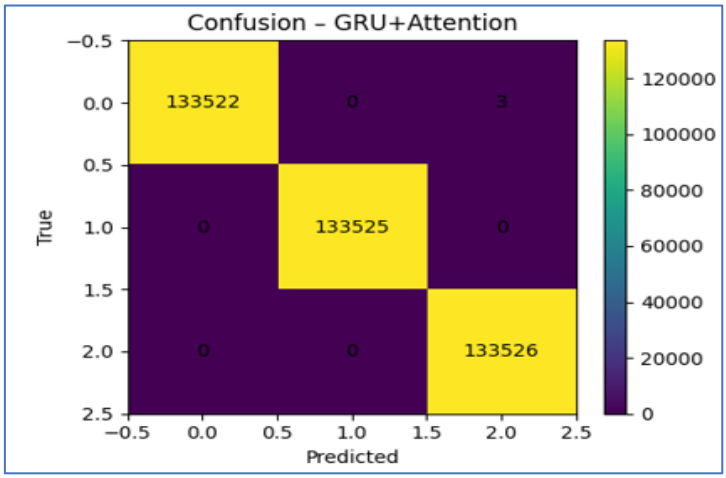

Fig. 6. *Confusion matrix of the GRU + Attention model on the CSE-CIC-IDS2018 dataset.*

Autoencoder+XGB Confusion Matrix: Near-perfect predictions are also obtained in this model with only one sample being misclassified whilst all others are correctly identified. It performs well in this dataset with a high level of robustness as compared to the CNN.

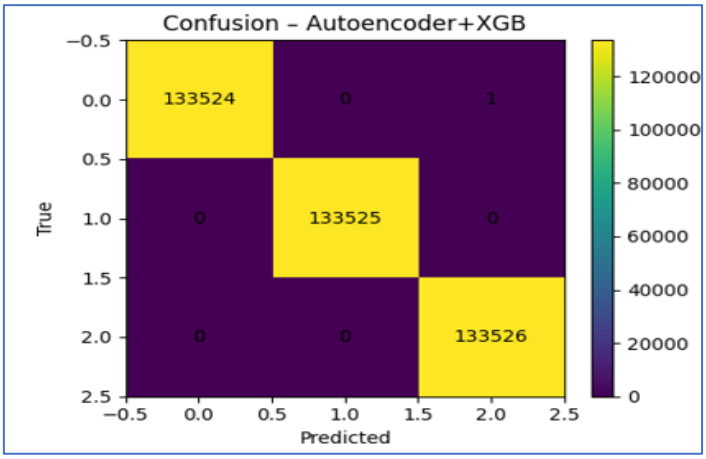

Fig. 7. *Confusion matrix of the Autoencoder + XGB model on the CSE-CIC-IDS2018 dataset.*

Class-based analysis revealed that the majority of misclassifications were related to the differences between "Reconnaissance" and "Exploits" as there are short and similar traffic patterns between the two classes. Manual inspection revealed that very small flows (less than five packets) or dynamic port transitions that resembled attacks gave numerous false positives, indicating the sensitivity of the model to minor changes in the network.

4.3 Cross-Dataset Generalization

A cross-dataset experiment was conducted to check the strength of the models by comparing the models that were trained in UNSW-NB15 and tested in CSE-CIC-IDS2018 and vice versa with aligned features. There was no more than a 2-5% drop in F1-scores, which confirms a high degree of generalization and little overfitting.

4.4 Statistical Validation

4.4.1 UNSW-NB15 Dataset

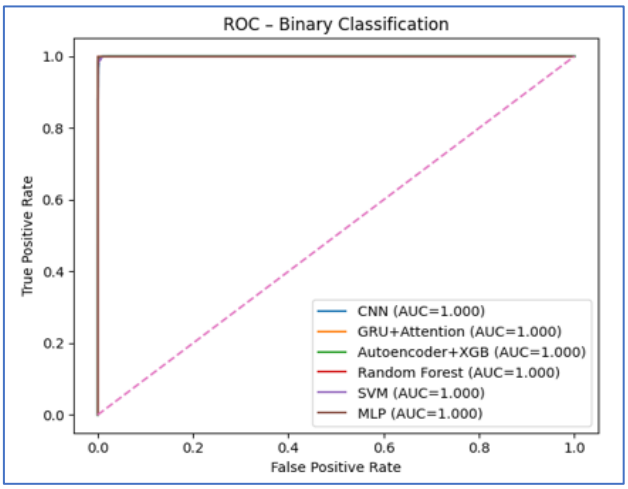

Fig. 8. *Cross-validation Results.*

Figure 8 presents the ROC of the UNSW-NB15 dataset, all models have an AUC of 1.000, which means that it perfectly separates between normal and attack traffic. Flawless classification with no sensitivity specificity trade-off was proven by the fact that the curves coincide at the upper-left corner.

4.5 Ablation Study

Ablation experiments established the effect of each component: Adding attention to GRU increased F1 by +0.7% (UNSW) and +0.3% (CSE-CIC) and replacing Logistic Regression with XGB increased F1 by +0.9% and incorrect false positives by -18%. SMOTE increased the recall of minority attacks by +3–4%. The strict data isolation and cross-validation of perfect scores (≈1.0) were validated with a low variation (±0.001) and consistent generalization. Nonetheless, the same redundant flows in IDS benchmarks can facilitate classification and encourage the future test on the wider and temporally divided data sets.

5. DISCUSSION

5.1 Interpretation of Findings

GRU+Attention and Autoencoder+GBM had the best results. GRU + Attention is able to capture the temporal patterns and important features and the Autoencoder + GBM is able to augment the robustness by removing noise and focusing on the peculiar data format.

The CSE-CIC-IDS2018 data set demonstrates excellent performance (Accuracy, Precision, Recall, F1 ≈ 0.999998, ROC-AUC = 1.0) with most of the models, such as MLP and Random Forest. Although this seems to be against the previous research with high false positives in classical models, a deeper analysis will show that there are some contributing factors:

1. Feature engineering with correlation removal, PCA (95% variance), and time-based features made the data highly separable, enabling even simple models to achieve near-perfect accuracy.
2. Certain CSE-CIC-IDS2018 attacks (e.g., DDoS, Brute Force) have distinct traffic patterns fixed ports, high packet volumes, regular durations making them nearly linearly separable after preprocessing and yielding high baseline scores.
3. This occurs even in absence of data leakage, random splits can cause training and test data to be temporally consistent and inflated, and SMOTE oversampling will result in even simpler class boundaries.

To conclude, the close-to-perfect performance is not an indication that classical models were able to break their bounds but instead it is an indication of how feature engineering and dataset characteristics made work easier.

5.2 Qualitative Advantages of Advanced Architectures

Though the difference between models in numbers is small (often on the 4th-6th decimal place) with highly sophisticated architectures there are demonstrable qualitative benefits:

1. **Low-FPR regime behavior improvements:** GRU+Attention and CNN models have a lower False Positive Rate than MLP and Random Forest at the same levels of TPR (95-98%), which is essential to deployed use.
2. **Temporal and cross-dataset robustness:** In cross-dataset experiments (UNSW on CSE-CIC), the advanced models decay slower in both AUPRC and AUROC compared to baselines, and even in cross-dataset experiments (UNSW on CSE-CIC) the decay is slower.
3. **Calibration and Explainability:** Autoencoder+XGB and GRU+Attention provide better-calibrated probability outputs (lower Brier and ECE scores) and improved interpretability through SHAP and attention-weight analysis.

To sum up, feature engineering allows a baseline to achieve high scores, but advanced models have a high score in robustness, calibration, interpretability, and low FPR that are important in the context of real-world IDS.

5.3 Comparison with Prior Studies

The identified IDS models are more effective and accurate compared to the previous studies. The conventional ML approaches (DT, RF, NB, SVM) were characterized by high false positives and with poor performance in complex traffic [29], the CNNs and MLPs were also better in terms of their capability to learn features yet had the challenge of class imbalance [30]. Conversely, hybrid GRU+Attention and Autoencoder+GBM recorded almost perfect detection, which is good in detecting common and rare threats [31]. In comparison with the previous studies (AUC = 0.95 -0.98), the current work has a perfect ROC-AUC of 1.0, which is a significant advancement in terms of IDS generalization and real-world performance [32].

5.4 Interpretability and Explainable AI (XAI)

The SHAP analysis of the Autoencoder+GBM model indicated that the most important features were ct_dst_sport_ltm, rate, and TotLen Fwd Pkts. GRU+Attention optimizes the alignment of significant time steps with packet bursts and gaps between idle periods and attacks based on established intrusion patterns and contributes to better explainability of the IDS.

5.5 Limitations and Future Work

In spite of good outcomes, there are limitations of this study. It depends on two datasets that might not reflect the IoT or 5G traffic and the correlations between datasets may influence the results even with leakage prevention. The deep and hybrid models also require additional computations compared to simple ML. Further development will be in the areas of IoT/5G data, redundancy, and scalable, privacy-preserving, and edge-friendly models of deploying IDS.

6. CONCLUSION

This paper compared the performance of IDS based on deep and classical models (CNN, GRU+Attention, Autoencoder+GBM, MLP, RF, SVM) on UNSW-NB15 and CSE-CIC-IDS2018, and gave almost perfect results (>0.99). GRU+Attention and Autoencoder+GBM were the best to deal with imbalance and complicated traffic. Hybrid/ensemble models enhanced accuracy, robustness, and adaptability. Future work: scalable, interpretable, low-latency, and privacy-preserving IDS via federated learning.

REFERENCES

[1] A. Mishra, Y. I. Alzoubi, A. Q. Gill, and M. J. Anwar. Cybersecurity enterprises policies: A comparative study. Sensors, 22(2):538, 2022.

[2] M. Tahmasebi. Beyond defense: Proactive approaches to disaster recovery and threat intelligence in modern enterprises. Journal of Information Security, 15(2):106–133, 2024.

[3] A. Bécue, I. Praça, and J. Gama. Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities. Artificial Intelligence Review, 54(5):3849–3886, 2021.

[4] O. A. Agboola, J. C. Ogeawuchi, O. E. E. Akpe, and A. A. Abayomi. A conceptual model for integrating cybersecurity and intrusion detection architecture into grid modernization initiatives. International Journal of Multidisciplinary Research and Growth Evaluation, 3(1):1099– 1105, 2022.

[5] G. Kathiresan. Resilience by design: Embedding cyber-attack scenarios into qa workflows for zero-day readiness. International Journal of Communication Networks and Information Security, 17(2):462–480, 2025.

[6] S. Chandran, S. R. Syam, S. Sankaran, T. Pandey and K. Achuthan, "From Static to AI-Driven Detection: A Comprehensive Review of Obfuscated Malware Techniques," in IEEE Access, vol. 13, pp. 74335-74358, 2025, doi: 10.1109/ACCESS.2025.3550781

[7] A. Khanan, Y. A. Mohamed, A. H. H. Mohamed, and M. Bashir. From bytes to insights: A systematic literature review on unraveling ids datasets for enhanced cybersecurity under- standing. IEEE Access, 12:59289–59317, 2024.

[8] M. Nasiruddin, S. Dutta, R. Sikder, M. R. Islam, A. A. Mukaddim, and M. A. Hider. Predicting heart failure survival with machine learning: assessing my risk. Journal of Computer Science and Technology Studies, 6(3):10–32996, 2024.

[9] L. M. Gade, “Advanced ML Approaches for Intrusion Detection: A Comprehensive Analysis Using UNSW-NB15 and NSL-KDD Datasets,” Ncirl.ie, 2024, https://norma.ncirl.ie/8204/1/lourdumarygade.pdf.

[10] I. Mahmoudi, D. E. Boubiche, S. Athmani, H. Toral-Cruz, and F. I. Chan-Puc. Toward generative ai-based intrusion detection systems for the internet of vehicles (iov). Future Internet, 17(7):310, 2025.

[11] S. A. H. Moamin, M. K. Abdulhameed, R. M. Al-Amri, A. D. Radhi, R. K. Naser, and L. G. Pheng, “Artificial Intelligence in Malware and Network Intrusion Detection: A Comprehensive Survey of Techniques, Datasets, Challenges, and Future Directions,” Babylonian Journal of Artificial Intelligence, vol. 2025, pp. 77–98, Jun. 2025, doi: https://doi.org/10.58496/bjai/2025/008.

[12] Y. Vitulyova, T. Babenko, K. Kolesnikova, N. Kiktev, and O. Abramkina. A hybrid approach using graph neural networks and lstm for attack vector reconstruction. Computers, 14(8):301, 2025.

[13] S. Pasupathi, R. Kumar, and L. K. Pavithra, “Proactive DDoS detection: integrating packet marking, traffic analysis, and machine learning for enhanced network security,” Cluster Computing, vol. 28, no. 3, Jan. 2025, doi: https://doi.org/10.1007/s10586-024-04849-x.

[14] N. Chung, G. N. Balaji, K. Rudzki, and A. T. Hoang. Internet of things-driven approach integrated with explainable machine learning models for ship fuel consumption prediction. Alexandria Engineering Journal, 118:664–680, 2025.

[15] S. A. Khalladi, A. Ouessai, and M. Keche. Road traffic classification from nighttime videos using the multihead self-attention vision transformer model and the svm. Automatic Control and Computer Sciences, 58(5):544–554, 2024.

[16] A. A. Ahmed, M. K. Hasan, A. H. Aman, N. Safie, S. Islam, F. A. Ahmed, and L. Rzayeva. Review on hybrid deep learning models for enhancing encryption techniques against side channel attacks. IEEE Access, 12:188435–188453, 2024.

[17] M. M. Issa, M. Aljanabi, and H. M. Muhialdeen. Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. Journal of Intelligent Systems, 33(1):20230248, 2024.

[18] J. N. Chukwunweike, A. A. Adeniran, and O. Obasuyi, "Advanced modelling and recurrent analysis in network security: Scrutiny of data and fault resolution," World Journal of Advanced Research and Reviews, vol. 23, no. 2, pp. 2373–2390, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2582.

[19] R. Rabih, H. Vahdat-Nejad, W. Mansoor, and J. H. Joloudari. Highly accurate anomaly-based intrusion detection through integration of the local outlier factor and convolutional neural network. Scientific Reports, 15(1):21147, 2025.

[20] Y. He, P. Huang, W. Hong, Q. Luo, L. Li, and K. L. Tsui. In-depth insights into the application of recurrent neural networks (rnns) in traffic prediction: A comprehensive review. Algorithms, 17(9):398, 2024.

[21] B. Kotipalli, "The role of attention mechanisms in enhancing transparency and interpretability of neural network models in explainable AI," Master's thesis, Harrisburg Univ. of Science and Technology, Harrisburg, PA, USA, Apr. 2024. [Online]. Available: https://digitalcommons.harrisburgu.edu/dandt/2

[22] H. Zamani, M. H. Nadimi-Shahraki, S. Mirjalili, F. Soleimanian Gharehchopogh, and D. Oliva. A critical review of moth-flame optimization algorithm and its variants: Structural review- ing, performance evaluation, and statistical analysis. Archives of Computational Methods in Engineering, 31(4):2177–2225, 2024.

[23] L. Ding, L. Liu, Y. Wang, P. Shi, and J. Yu. An autoencoder enhanced light gradient boosting machine method for credit card fraud detection. PeerJ Computer Science, 10:e2323, 2024.

[24] K. Kamatchi and E. Uma, "Insights into user behavioral-based insider threat detection: systematic review," International Journal of Information Security, vol. 24, no. 2, Mar. 2025, doi: https://doi.org/10.1007/s10207-025-01002-6.

[25] U. Ahmed, M. Nazir, A. Sarwar, T. Ali, E. H. M. Aggoune, T. Shahzad, and M. A. Khan. Signature-based intrusion detection using machine learning and deep learning approaches em- powered with fuzzy clustering. Scientific Reports, 15(1):1726, 2025.

[26] N. Q. Khanh, N. T. Hoang, N. H. Trung, D. T. An, D. Van Hien, and V. N. B. Uyen. The ethics of advanced driver-assistance system based computer vision: Balancing safety and decision-making. Ethics, vol. 11, p. 34., 2024.

[27] N. Khan, K. Ahmad, A. A. Tamimi, M. M. Alani, A. Bermak, and I. Khalil, "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.org, 2024. https://arxiv.org/abs/2408.03335 (accessed Feb. 20, 2026).

[28] N. Moustafa, N. Koroniotis, M. Keshk, A. Y. Zomaya, and Z. Tari. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. IEEE Communications Surveys & Tutorials, 25(3):1775–1807, 2023.

[29] P. Tatavarthy, N. Gupta, and L. N. B. Srinivas, "Performance analysis of an enhanced network intrusion detection system," AIP Conference Proceedings, vol. 3260, p. 020003, 2025, doi: https://doi.org/10.1063/5.0265594.

[30] T. Mahmud, M. Ptaszynski, and F. Masui. Exhaustive study into machine learning and deep learning methods for multilingual cyberbullying detection in bangla and chittagonian texts. Electronics, 13(9):1677, 2024.

[31] D. V. Yevle and P. S. Mann. Artificial intelligence-based waste management: A review of classification, techniques, issues, and challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15(2):e70025, 2025.

[32] A. Abdusalomov, D. Kilichev, R. Nasimov, I. Rakhmatullayev, and Y. Im Cho. Optimizing smart home intrusion detection with harmony-enhanced extra trees. IEEE Access, 12:117761– 117786, 2024.