Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

AI-Enhanced Intrusion Detection: Integrating Expert Knowledge and Machine Learning for Enterprise Networks

Fatima Zohra Allam*

*Corresponding author, Assistant Prof., Signal and Communication Laboratory, Department of Electronics, National Polytechnic School, Algeria. E-mail: fatima_zohra.allam@g.enp.edu.dz

Hicham Bousbia-Salah

Prof., Signal and Communication Laboratory, Department of Electronics, National Polytechnic School, Algeria. E-mail: hicham.bousbia-salah@g.enp.edu.dz

Fairouz Zendaoui

Assistant Prof., Laboratoire de la Communication dans les Systèmes Informatiques, Ecole Nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. E-mail: f_zendaoui@esi.dz

Latifa Hamami-Mitiche

Prof., Signal and Communication Laboratory, Department of Electronics, National Polytechnic School, Algeria. E-mail: latifa.hamami@g.enp.edu.dz

Journal of Information Technology Management, 2025, Vol. 17, Issue 4, pp. 99-116
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2025.105485
Article Type: Research Paper

Received: June 02, 2025
Received in revised form: August 16, 2025
Accepted: September 19, 2025
Published online: November 01, 2025

Abstract

Enterprise networks, as the backbone of modern information systems, are increasingly exposed to sophisticated and rapidly evolving cyber threats. Traditional Intrusion Detection Systems (IDS), based on static attack signatures, often fail to detect novel or complex intrusions, resulting in high false alarm rates. This study proposes an intelligent IDS that leverages Machine Learning and Deep Learning techniques to significantly improve detection accuracy and reduce alert noise. The system is capable of classifying attacks by severity and provides an intuitive interface to support efficient threat monitoring. Beyond technical performance, the solution addresses managerial objectives by lowering maintenance costs, enhancing service quality, accelerating incident response, and ensuring high availability with

straightforward deployment. The proposed model offers a scalable and resilient IDS tailored for enterprise environments, contributing both practical and strategic value in the fight against increasingly sophisticated cyberattacks.

Keywords: Intrusion Detection Systems, Artificial Intelligence, Deep Learning, Machine Learning

Introduction

Enterprise networks form the backbone of modern Information Systems (IS), enabling the continuous expansion of digital technologies and the transmission of increasingly diverse and complex data. While this rapid evolution provides greater functionality, it also broadens the attack surface and heightens vulnerabilities to security breaches that exploit architectural weaknesses.

In this context, a major challenge for network administrators is to strengthen prevention, detection, and response capabilities against cyberattacks. Intrusion Detection Systems (IDS) are central to this defense, as they monitor network activities and identify suspicious or malicious behaviors. However, traditional IDSs are predominantly signature-based, limiting their ability to detect unknown or sophisticated threats. Their rigidity often results in high false positive and false negative rates, thereby reducing overall effectiveness.

To overcome these limitations, Artificial Intelligence (AI) has emerged as a promising solution. By leveraging Machine Learning and Deep Learning, IDS can learn from large-scale datasets, recognize complex attack patterns, and adapt to novel threats. Such intelligent systems enhance detection accuracy, reduce alert fatigue, and provide meaningful classification of attacks based on type and severity.

Building on this perspective, our research aims to design and implement an intelligent IDS that integrates advanced AI techniques. The objectives are threefold:

- Improve intrusion detection performance in enterprise environments;
- Develop a system capable of classifying attacks by type and severity;
- Provide an intuitive interface to facilitate alert interpretation.

This study contributes to the broader effort of strengthening cybersecurity by combining technical robustness with practical usability. Our approach offers a proactive and adaptive solution aligned with enterprise security requirements, resource optimization, and agile threat response.

To structure the study, Section 2 reviews the main categories of cyberattacks. Section 3 introduces fundamental concepts of information system protection, with a focus on IDS

principles and classifications. Section 4 presents the technical implementation of our models and evaluates their performance using standard metrics (precision, recall, F-measure, etc.). Section 5 illustrates the model construction and evaluation. Finally, Section 6 concludes the work, highlights limitations, and outlines directions for future research to further optimize AI-driven intrusion detection.

Main Categories of Network Attack

In today's rapidly evolving digital landscape, enterprise networks face an increasingly complex and dynamic threat environment. Cyberattacks no longer consist of isolated incidents but often represent coordinated, multi-vector campaigns exploiting both technical vulnerabilities and human factors. An attack can be broadly defined as any malicious action intended to compromise one or more fundamental security properties: confidentiality, integrity, or availability (Hasan et al., 2023; Djuitcheu et al., 2022).

Figure 1 provides an overview of the principal categories of attacks most frequently encountered in enterprise environments.

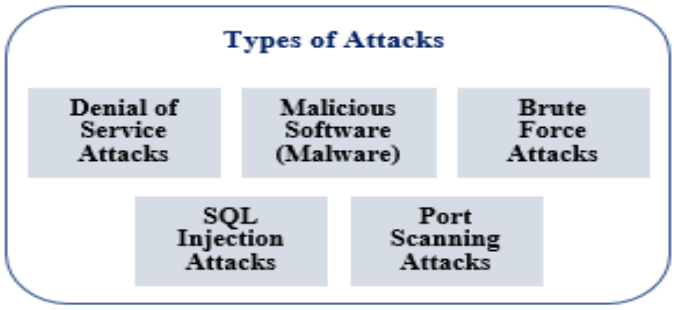


Figure 1. Types of Attacks

A critical review of these threats shows that while their basic forms are well known, attackers continuously adapt them to evade traditional detection methods. This evolutionary dimension underscores the need for advanced intrusion detection systems that integrate both expert knowledge and machine learning capabilities.

Denial of Service (DoS) and Distributed Denial of Service (DDoS) Attacks

DoS and DDoS attacks remain among the most disruptive threats because of their simplicity to launch and their potentially devastating impact. They aim to exhaust system resources and render services unavailable to legitimate users (Rustam et al., 2022; Chaganti et al., 2022).

- **Smurf Attack**: Exploits ICMP packets and broadcast amplification to overwhelm the victim with traffic (Revathy et al., 2022; Bouyeddou et al., 2018). While traditional defenses filter broadcast traffic, new amplification vectors (e.g., DNS, NTP) have emerged, showing the persistence of this threat.

- **SYN Flooding**: Abuses the TCP three-way handshake by sending numerous SYN requests without completing the connection, depleting server resources (Zeebaree et al., 2020). Despite long-standing mitigation strategies, attackers still combine SYN floods with other techniques to bypass detection.
- **DDoS**: Utilizes botnets, often built from IoT devices with poor security practices, to launch large-scale, coordinated service disruptions (Rustam et al., 2022). The sheer volume of such attacks challenges both signature-based detection and anomaly detection systems.

Malware (Malicious Software)

Malware represents a continuously evolving threat landscape. It refers to software specifically designed to infiltrate, disrupt, or damage computer systems. Common forms include worms, trojans, and ransomware, which pursue objectives ranging from data theft to complete system takeover (Chaganti et al., 2022).

A critical aspect is that malware is no longer static; it increasingly leverages polymorphism and fileless execution techniques to evade signature-based defenses. This adaptability highlights the necessity of detection models capable of recognizing malicious behavior patterns rather than relying solely on static features.

Brute-Force Attacks

Brute-force attacks involve systematically attempting all possible combinations of passwords or encryption keys to obtain unauthorized access. These attacks are particularly effective against systems with weak authentication mechanisms (Najafabadi et al., 2014). While theoretically simple, their success often reflects persistent organizational weaknesses, such as inadequate password policies or delayed adoption of multi-factor authentication.

From a detection standpoint, differentiating Brute-Force attempts from legitimate login errors presents a non-trivial challenge, especially in large-scale enterprise systems.

SQL Injection (SQLi) Attacks

SQL injection exploits poorly secured input fields by inserting malicious SQL statements. This allows attackers to access, manipulate, or delete sensitive database content (Roy et al., 2022). Despite being a well-documented and preventable vulnerability, SQLi continues to rank among the most critical threats.

Its persistence indicates the gap between secure development practices and real-world implementation, underlining the importance of proactive monitoring and automated detection mechanisms.

Port Scanning

Port scanning is a reconnaissance technique used to detect open ports and running services on a target system. While not inherently malicious, it is often a precursor to more targeted attacks (Aamir et al., 2021). The critical challenge lies in distinguishing benign scanning (e.g., vulnerability assessments) from adversarial probing.

This ambiguity necessitates contextual analysis, an area where expert-driven rules combined with machine learning approaches can significantly enhance accuracy.

Table 1 summarizes the main categories of attacks commonly affecting enterprise environments.

Table 1. Main Categories of Attacks

Attack Type	Objective	Technique	Impact
Denial of Service (DoS / DDoS)	Render a service unavailable	Resource exhaustion via ICMP floods (Smurf), TCP handshake abuse (SYN Flood), or botnet-based large-scale traffic (DDoS)	Service disruption, loss of availability, degraded user experience
Malware (Worms, Trojans, Ransomware)	Infiltrate, disrupt, or control systems	Malicious code execution, file encryption, unauthorized access	Data theft, system damage, ransom demands, persistent compromise
Brute-Force Attacks	Gain unauthorized access	Systematic trial of password or encryption key combinations	Compromised accounts, unauthorized privilege escalation
SQL Injection (SQLi)	Exploit database vulnerabilities	Insertion of malicious SQL commands into input fields	Unauthorized data access, modification, or deletion
Port Scanning	Reconnaissance before an attack	Probing for open ports and running services	Identification of system weaknesses enabling targeted intrusions

Intrusion Detection Systems (IDS)

Principles and Definitions

With the growing sophistication of cyber threats, organizations rely on multiple layers of defense, such as access control, cryptographic protocols, antivirus software, firewalls, and regular security audits. While these mechanisms form an essential first line of defense, they are insufficient in isolation to provide comprehensive protection. Attackers continuously develop techniques to bypass preventive measures, thereby exposing systems to persistent risks. To address this gap, Intrusion Detection Systems (IDS) were introduced as a complementary second line of defense (Labonne, 2020).

An IDS can be defined as a hardware-, software-, or hybrid-based security solution designed to monitor, analyze, and detect suspicious or unauthorized activities in real time. Its objectives include identifying intrusion attempts, detecting malware activity, and recognizing

abnormal network traffic patterns that may indicate ongoing attacks. By generating timely alerts, IDS assists administrators in initiating rapid and effective responses.

Typically, an IDS comprises three fundamental components (El Rab, 2008; Lindstedt, 2022):

- Sensor (Probe): Collects relevant data (e.g., network packets, system logs, or system calls) and forwards it to the analysis engine.
- Analyzer (Engine): Acts as the core component, processing incoming data to detect potential intrusions. If malicious activity is suspected, the analyzer triggers an alert.
- Response Module (Response Manager): Manages alerts and initiates appropriate countermeasures. Responses may be:
- Passive: Logging the event or notifying the administrator.
- Active: Automatically disrupting connections, resetting sessions, or dynamically updating firewall rules.

Through this layered structure, IDS contributes to situational awareness and proactive defense, complementing traditional preventive mechanisms.

IDS Classification

IDS can be classified along two primary dimensions: the location of the monitored data source and the detection methodology employed.

Based on the Data Source

- Host-based IDS (HIDS): Deployed on individual endpoints, HIDS monitors system calls, file integrity, and log activity. They are effective against localized threats such as Trojans and can analyze encrypted traffic, though they may impact host performance and scalability.
- Network-based IDS (NIDS): Positioned within the network, NIDS inspects traffic in real time to detect suspicious patterns or known attack signatures. While they provide rapid, system-wide monitoring without burdening hosts, their efficiency declines in high-speed or encrypted traffic environments.
- Hybrid IDS: Combine the strengths of HIDS and NIDS, offering more comprehensive visibility but at the cost of greater deployment complexity and resource requirements.

Based on the Detection Method

- Signature-Based Detection: Compares observed activity with a database of known attack patterns. It is fast and reliable for well-documented threats but incapable of detecting novel or zero-day attacks.

- Anomaly-Based Detection: Builds a model of normal system or network behavior and flags deviations as potential threats. While suitable for detecting unknown or evolving attacks, it suffers from high false positive rates, especially in dynamic environments.

Limitations of Traditional IDS

Despite their central role in cybersecurity, IDSs face persistent challenges that limit their effectiveness in enterprise contexts:

- Resource Overhead: HIDS are computationally demanding and themselves vulnerable to denial-of-service (DoS) conditions.

- Scalability Issues: NIDS may struggle to process traffic in high-bandwidth networks or under heavy load, leading to missed detections.

- Accuracy Concerns: Anomaly-based systems are prone to generating large volumes of false positives due to imperfect behavioral baselines, overwhelming administrators.

- Detection Latency: Many IDS are not fully capable of real-time analysis, delaying response and allowing attackers to exploit vulnerabilities.

- Adaptability Gaps: Static signature-based approaches cannot keep pace with polymorphic malware, encrypted traffic, and zero-day exploits.

These limitations highlight the urgent need for intelligent, adaptive, and resource-efficient IDS. By integrating expert-driven rules with advanced Artificial Intelligence techniques, particularly Machine Learning and Deep Learning, next-generation IDS can enhance detection accuracy, reduce alert fatigue, and provide more actionable insights for enterprise security management.

Implementation and Testing

Artificial Intelligence (AI) has emerged as a powerful enabler for enhancing both the accuracy and adaptability of Intrusion Detection Systems (IDS). In this work, we designed and implemented a hybrid approach that integrates two complementary components:

- A Deep Learning (DL) architecture, primarily aimed at learning baseline network behavior and detecting deviations indicative of anomalies. This component directly

addresses the core challenge of IDS: distinguishing between benign and malicious activity in dynamic and complex environments.

- A Machine Learning (ML) architecture, specifically tailored for the fine-grained classification of detected intrusions, enabling a detailed identification of attack types and their severity.

Figure 2 provides an overview of the proposed solution, highlighting the interplay between the DL-based anomaly detection module and the ML-based classification module.

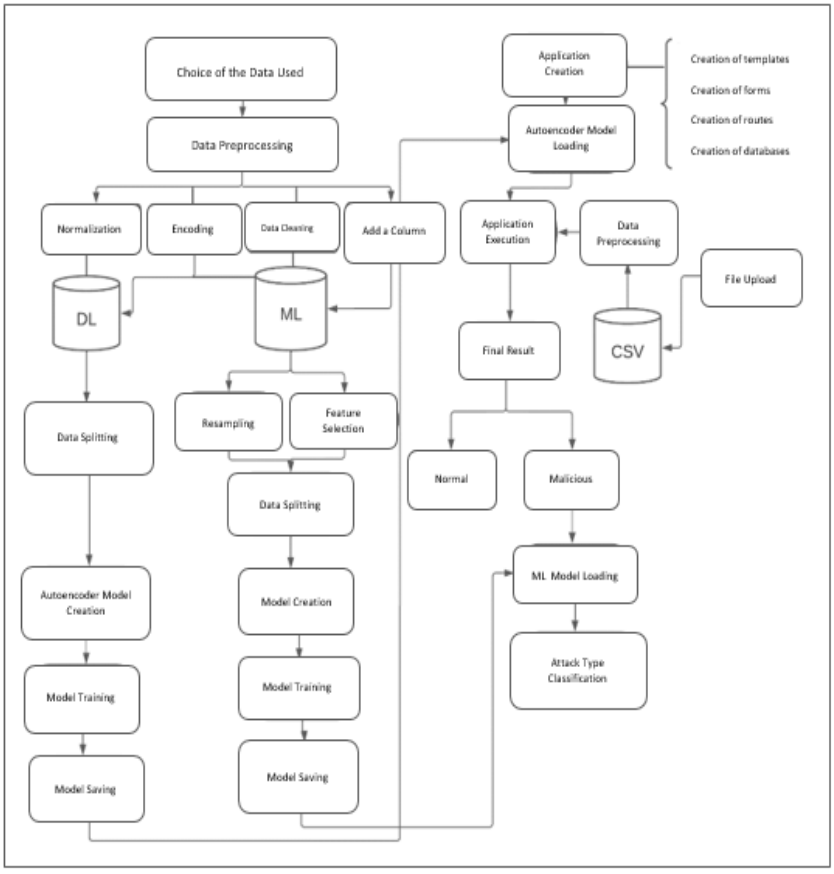
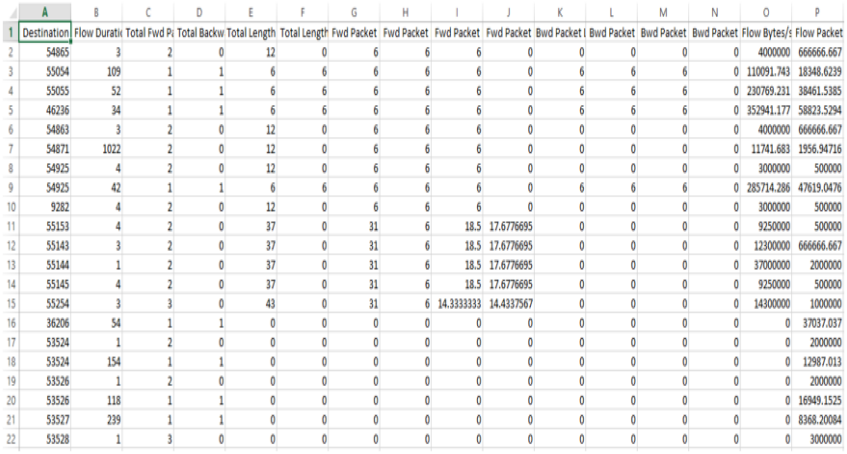


Figure 2. Overview of the Solution

To evaluate our approach, we relied on the CICIDS-2017 dataset, provided by the Canadian Institute for Cybersecurity (ISCX Consortium). This dataset was selected due to its close alignment with real-world enterprise traffic conditions. It comprises eight network monitoring sessions, stored in CSV format, with a total of 2.829.385 records across 79 attributes. Figure 3 presents an excerpt of the dataset format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Destination	Flow Duratio	Total Fwd Pa	Total Backw	Total Length	Total Length	Fwd Packet	Fwd Packet	Fwd Packet	Fwd Packet	Bwd Packet L	Bwd Packet	Bwd Packet	Bwd Packet	Flow Bytes/s	Flow Packet
2	54865	3	2	0	12	0	6	6	6	0	0	0	0	0	4000000	666666.667
3	55054	109	1	1	6	6	6	6	6	0	6	6	6	0	110091.743	18348.6239
4	55055	52	1	1	6	6	6	6	6	0	6	6	6	0	230769.231	38461.5385
5	46236	34	1	1	6	6	6	6	6	0	6	6	6	0	352941.177	58823.5294
6	54863	3	2	0	12	0	6	6	6	0	0	0	0	0	4000000	666666.667
7	54871	1022	2	0	12	0	6	6	6	0	0	0	0	0	11741.683	1956.94716
8	54925	4	2	0	12	0	6	6	6	0	0	0	0	0	3000000	500000
9	54925	42	1	1	6	6	6	6	6	0	6	6	6	0	285714.286	47619.0476
10	9282	4	2	0	12	0	6	6	6	0	0	0	0	0	3000000	500000
11	55153	4	2	0	37	0	31	6	18.5	17.6776695	0	0	0	0	9250000	500000
12	55143	3	2	0	37	0	31	6	18.5	17.6776695	0	0	0	0	12300000	666666.667
13	55144	1	2	0	37	0	31	6	18.5	17.6776695	0	0	0	0	37000000	2000000
14	55145	4	2	0	37	0	31	6	18.5	17.6776695	0	0	0	0	9250000	500000
15	55254	3	3	0	43	0	31	6	14.3333333	14.4337567	0	0	0	0	14300000	1000000
16	36206	54	1	1	0	0	0	0	0	0	0	0	0	0	0	37037.037
17	53524	1	2	0	0	0	0	0	0	0	0	0	0	0	0	2000000
18	53524	154	1	1	0	0	0	0	0	0	0	0	0	0	0	12987.013
19	53526	1	2	0	0	0	0	0	0	0	0	0	0	0	0	2000000
20	53526	118	1	1	0	0	0	0	0	0	0	0	0	0	0	16949.1525
21	53527	239	1	1	0	0	0	0	0	0	0	0	0	0	0	8368.20084
22	53528	1	3	0	0	0	0	0	0	0	0	0	0	0	0	3000000

Figure 3. Dataset Format

The CICIDS-2017 dataset is particularly suited for IDS research as it combines:

- Realistic scale and diversity, including over two million benign records alongside numerous attack types;
- Rich feature space, capturing diverse traffic characteristics relevant to modern network architectures;
- Comprehensive coverage of attack scenarios, ranging from volumetric attacks to sophisticated application-layer exploits

This dataset thus provides a robust experimental foundation for training, validating, and benchmarking our intelligent IDS.

Data Preprocessing

The quality and relevance of training data are decisive factors for the performance of AI-based IDS. To this end, we applied a rigorous preprocessing pipeline designed to improve data quality, ensure comparability across features, and mitigate dataset biases.

Data Cleaning

Extensive data cleaning was performed to eliminate inconsistencies and noise. These anomalies, which could affect the results, are removed to ensure the model's reliability. This included eliminating:

- Missing values and duplicate rows,
- Columns with only null values,
- Infinite values,
- Statistically identified outliers.

Data Encoding

Since ML/DL models require numerical inputs, categorical labels were converted into integer values using the Label-Encoder function. The "Label" column, originally containing attack categories in text form, was thus encoded into a machine-readable format.

Data Normalization

We applied z-score normalization to all numerical features, excluding "Label" and "Quality", ensuring standardized distributions across attributes. This step was crucial for stabilizing learning processes, particularly for algorithms sensitive to scale variations.

Feature Selection

For the ML-based classification model, we reduced feature redundancy through a supervised feature selection process. Feature importance was visualized (Figure 4), and attributes below a predefined threshold were removed. Additionally, highly correlated variables were eliminated following correlation matrix analysis. This optimization improved both computational efficiency and model interpretability.

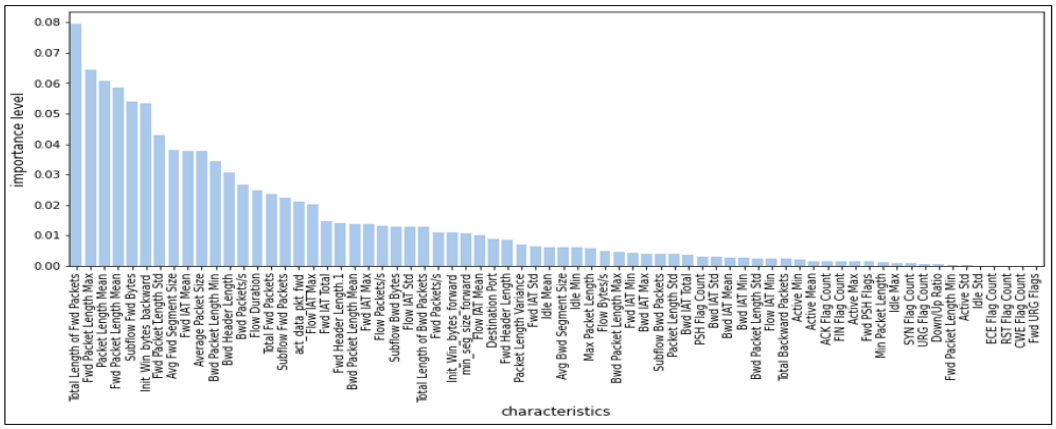


Figure 4. Overview of Important Features

Resampling

Analysis of the distribution of attack types within the dataset, shown in Figure 5, revealed a significant imbalance between classes (unbalanced data). This imbalance is a major challenge in the CICIDS-2017 dataset, which can significantly impair the performance of learning models, particularly for minority class detection.

To address this issue, we applied the Synthetic Minority Oversampling Technique (SMOTE) algorithm, which generates synthetic samples of minority classes by interpolating between existing instances. The resulting balanced distribution is shown in Figure 6, demonstrating better representativeness of minority attacks.

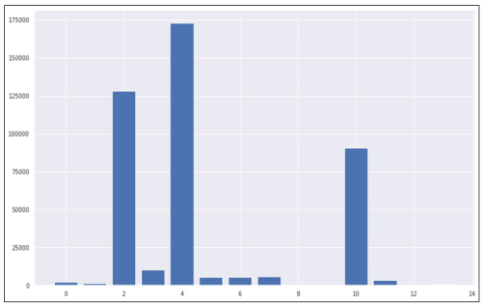


Figure 5. Attack-Type Values Before Oversampling

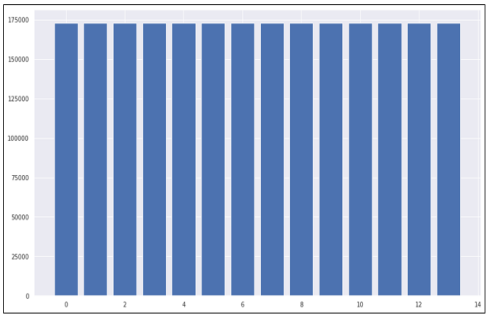


Figure 6. Attack-Type Values After Oversampling

Data Splitting

For effective model training and unbiased evaluation, we carefully structured the dataset. The "BENIGN" class was preserved as a key reference for distinguishing legitimate from malicious traffic. To operationalize this, we derived a binary column separating normal versus malicious flows.

We then applied a standard Train-Test split (80/20), ensuring representative distributions of both benign and attack traffic. Table 2 summarizes the partitioned dataset, which served as the basis for training and performance evaluation of the DL and ML models.

Table 2. Distribution of Data

Model	Train	Test	Total
DL	1676045	419012	2095057
ML	340592	85149	425741

Model Construction and Evaluation

Autoencoder Model

To address the challenge of detecting anomalies in network traffic, we first developed an Autoencoder architecture, which is particularly suited for unsupervised intrusion detection. The intuition is that by training on normal traffic, the Autoencoder learns a compressed representation of legitimate behavior and consequently struggles to reconstruct malicious inputs, leading to higher reconstruction errors.

The model is composed of three parts: the input layer, encoder layers, and decoder layers. The architecture, specifically the number of layers, neurons per layer, and activation functions, was selected empirically after experimenting with several configurations. The retained structure provided the most stable convergence and the lowest reconstruction error.

Once the architecture was fixed, we compiled the model with the following parameters:

- Optimizer: Adam, chosen for its adaptive learning rate and efficiency in handling sparse

gradients, which is particularly useful in high-dimensional datasets like network traffic.

- Loss Function: Mean Absolute Error (MAE), as it is less sensitive to outliers than Mean Squared Error (MSE) and aligns well with anomaly detection goals.
- Epochs: 20, representing a balance between sufficient learning and the avoidance of overfitting.
- Batch size: 64, enabling efficient training while maintaining generalization capability.

The model training process is depicted in Figure 7, showing the progressive reduction of reconstruction error.

```
- loss: 0.1979 - accuracy: 0.9351 - val_loss: 0.1980 - val_accuracy: 0.9135
- loss: 0.1964 - accuracy: 0.9362 - val_loss: 0.1987 - val_accuracy: 0.9212
- loss: 0.1959 - accuracy: 0.9366 - val_loss: 0.1969 - val_accuracy: 0.9395
- loss: 0.1961 - accuracy: 0.9370 - val_loss: 0.1959 - val_accuracy: 0.9445
- loss: 0.1967 - accuracy: 0.9358 - val_loss: 0.1973 - val_accuracy: 0.9595
- loss: 0.1965 - accuracy: 0.9359 - val_loss: 0.1975 - val_accuracy: 0.9662
- loss: 0.1963 - accuracy: 0.9370 - val_loss: 0.1959 - val_accuracy: 0.9721
```

Figure 7. Training of the Autoencoder

Visualization of Loss Results

The evolution of training and testing loss is illustrated in Figure 8. The gradual convergence of the curves reflects stable learning and the absence of severe overfitting. Nonetheless, the relatively small gap between training and validation errors does not fully guarantee robustness, particularly under adversarial or previously unseen traffic patterns. Incorporating additional validation strategies, such as cross-validation or evaluation on temporally separated datasets, would provide stronger evidence of the model's long-term reliability.

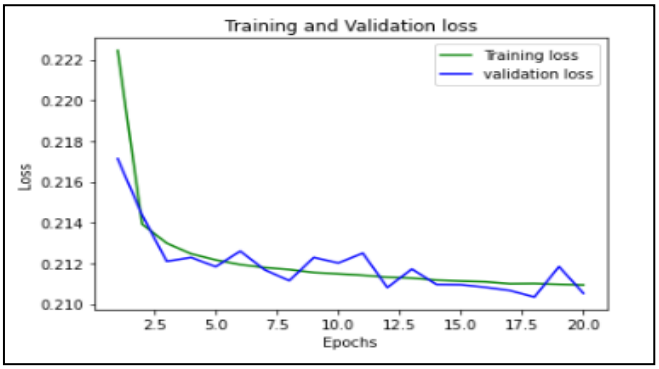


Figure 8. Visualization of Loss Results

Attack Type Classification Model

To complement the unsupervised Autoencoder, we implemented a supervised classification framework. Several widely adopted algorithms, Logistic Regression, Gradient Boosting, Naïve Bayes, and Random Forest, were evaluated in terms of Accuracy and Loss (see Table 3).

Table 3. Comparison between Performance Metrics

Different Algorithms / Performance Indicators	Logistic Regression	Gradient Boosting	Naïve Bayes	Random Forest
Accuracy	0.992	0.960	0.807	0.999
Loss	0.620	0.803	1.252	0.158

While most models achieved high accuracy, the Random Forest algorithm outperformed others, achieving near-perfect accuracy (0.999) and the lowest loss value (0.158). However, such near-optimal results warrant caution. They could indicate an excellent fit to the dataset but also raise the possibility of overfitting, particularly if the test set is not fully representative of real-world conditions.

The accuracy/loss comparison is further visualized in Figure 9, emphasizing Random Forest's dominance. Still, it is important to note that other metrics, such as precision, recall, and F1-score, are equally critical in intrusion detection, as high overall accuracy can mask poor detection of minority attack classes.

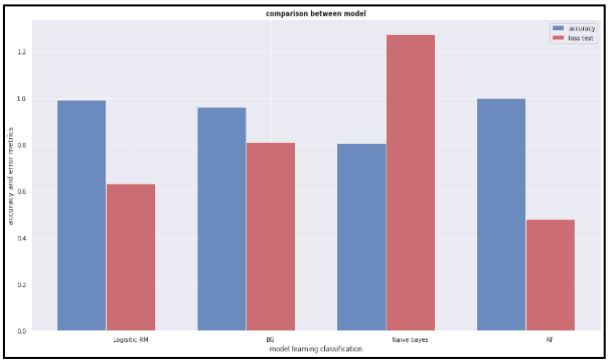


Figure 9. Classification Algorithms Results

Results

The confusion matrix of the Random Forest classifier (Figure 10) reveals its strong differentiation capability between benign and malicious traffic.

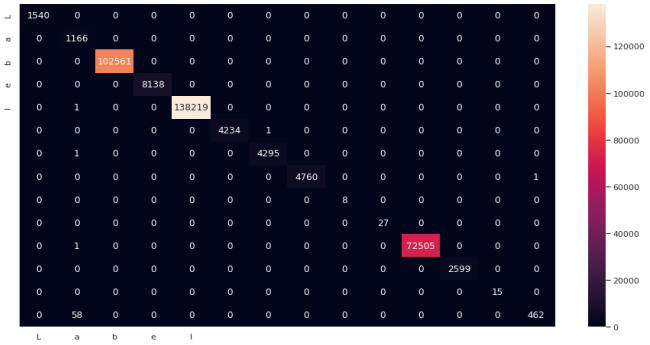


Figure 10. Confusion Matrix

From this matrix, we computed detailed performance indicators (precision, recall, and F1-score) for each attack class, presented in Figure 11. These results confirm the model's robustness, with consistently high values across all classes.

Classification report Training

	precision	recall	f1-score	support
Bot	1.00	1.00	1.00	1540
Brute Force	0.95	1.00	0.97	1166
DDoS	1.00	1.00	1.00	102561
DoS GoldenEye	1.00	1.00	1.00	8138
DoS Hulk	1.00	1.00	1.00	138220
DoS Slowhttptest	1.00	1.00	1.00	4235
DoS slowloris	1.00	1.00	1.00	4296
FTP-Patator	1.00	1.00	1.00	4761
Heartbleed	1.00	1.00	1.00	8
Infiltration	1.00	1.00	1.00	27
PortScan	1.00	1.00	1.00	72506
SSH-Patator	1.00	1.00	1.00	2599
Sql Injection	1.00	1.00	1.00	15
XSS	1.00	0.89	0.94	520

Figure 11. Performance Indicators

Nevertheless, several limitations should be acknowledged:

- Dataset dependency: The evaluation was conducted on a single dataset. Generalization to other network environments or real-world traffic remains untested.
- Feature interpretability: While Random Forest provides variable importance scores, the black-box nature of the model makes it challenging to extract actionable insights about the underlying attack mechanisms.
- Class imbalance: Some attack categories may be underrepresented, which could inflate the classifier's apparent performance if not carefully accounted for.

To enable reproducibility and future deployment, the trained Random Forest model was serialized using the pickle library in ".sav" format.

Synthesis

In summary, we implemented and developed two complementary approaches:

- An Autoencoder for unsupervised anomaly detection, which effectively learns the structure of normal traffic but whose performance could benefit from more systematic hyperparameter optimization.
- A supervised classifier, where Random Forest emerged as the most effective algorithm, achieving near-perfect classification performance, albeit with some cautionary notes regarding overfitting and dataset dependency.

Taken together, these results highlight the strength of hybrid approaches in intrusion detection: unsupervised learning captures deviations from normal behavior, while supervised classification ensures precise attack categorization. However, further validation across heterogeneous datasets and adversarial scenarios is essential before deploying such models in production-level systems.

Conclusion

Modern enterprise networks, marked by increasing heterogeneity and complexity, present expanding attack surfaces vulnerable to advanced cyber threats. This evolving landscape has intensified research efforts to fortify Intrusion Detection Systems (IDS). Despite significant advances, current IDS solutions still grapple with persistent challenges, including high false positive rates, extensive preprocessing needs, and limited resilience to novel or zero-day threats, highlighting a persistent gap between controlled experimental performance and practical deployment.

In this study, we addressed these challenges by developing two AI-based components: a deep learning–based autoencoder for unsupervised anomaly detection grounded in learned representations of normal network behavior, and a supervised classification model for precise attack-type identification. Our experiments demonstrate that this hybrid architecture can significantly enhance detection accuracy and classification precision, showcasing the potential of AI to elevate IDS toward more adaptive, context-aware defense systems.

However, the limitations are significant and must be acknowledged. One key issue is explainability: Deep Learning models often act as "black boxes", a concern well-documented in recent reviews on explainable IDS (X-IDS) and their sociotechnical implications (Neupane et al., 2022). Equally crucial is adversarial robustness. ML-based IDS remains vulnerable to carefully crafted evasion or poisoning attacks, underscoring the necessity for more resilient models.

Our reliance on a single dataset (CICIDS-2017) also limits generalization. Comprehensive surveys emphasize the need to validate IDS across diverse network environments, including IoT and SDN contexts, to ensure real-world applicability (Kumar et al., 2025). Furthermore, the gap between benchmarking studies and operational deployment,

particularly concerning latency, real-time processing, and system scalability, remains a critical obstacle (Kumar et al., 2025).

Nevertheless, several key directions remain open. We propose some future directions (Khan et al., 2024; Kumar et al., 2025; Wang et al., 2023):

- Cross-dataset validation: Train and evaluate models on multiple heterogeneous datasets to ensure robustness across varied network environments.
- Explainability: Integrate X-IDS methods, such as SHAP or LIME, enabling transparent decision-making and facilitating analyst trust and actionable insights.
- Adversarial Resilience: Incorporate defense mechanisms against evasion and poisoning, informed by emerging research on adversarial threats in ML-based IDS.
- Real-time performance optimization: Adapt architectures for low-latency deployment, particularly relevant in high-speed and resource-constrained environments like IoT or cloud-edge ecosystems.
- Scalable deployment strategies: Explore federated, edge, or hybrid learning frameworks to enable distributed IDS without compromising privacy or performance.

In conclusion, while our study presents promising evidence that AI-driven hybrids can outperform traditional IDS, it also underscores that such solutions are necessary yet insufficient on their own. Bridging the gap from promising prototypes to reliable production systems requires addressing explainability, adversarial resilience, deployment scalability, and cross-domain robustness. This paper lays a foundation, not the finish line, for building a trustworthy and adaptive IDS capable of proactively defending against an ever-evolving cyber threat landscape.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aamir, M., Rizvi, S. S. H., Hashmani, M. A., Zubair, M., & Ahmad, J. (2021). Machine learning classification of port scanning and DDoS attacks: A comparative analysis. *Mehran University Research Journal of Engineering & Technology*, *40*(1), 215-229.

Bouyeddou, B., Harrou, F., Sun, Y., & Kadri, B. (2018, May). Detection of smurf flooding attacks using a Kullback-Leibler-based scheme. In *2018 4th International Conference on Computer and Technology Applications (ICCTA)* (pp. 11-15). IEEE.

Chaganti, R., Boppana, R. V., Ravi, V., Munir, K., Almutairi, M., Rustam, F., ... & Ashraf, I. (2022). A comprehensive review of denial of service attacks in the blockchain ecosystem and open challenges. *IEEE Access*, *10*, 96538-96555.

Djuitcheu, H., Debes, M., Aumüller, M., & Seitz, J. (2022, March). Recent review of distributed denial of service attacks in the Internet of Things. In *2022 5th conference on cloud and internet of things (CIoT)* (pp. 32-39). IEEE.

El Rab, M. G. (2008). *Evaluation des systèmes de détection d'intrusion* (Doctoral dissertation, Université Paul Sabatier-Toulouse III).

Hasan, M. K., Habib, A. A., Islam, S., Safie, N., Abdullah, S. N. H. S., & Pandey, B. (2023). DDoS: Distributed denial of service attack in communication standard vulnerabilities in smart grid applications and cybersecurity with recent developments. *Energy Reports*, *9*, 1318-1326.

Khan, N., Ahmad, K., Tamimi, A. A., Alani, M. M., Bermak, A., & Khalil, I. (2024). Explainable AI-based Intrusion Detection System for Industry 5.0: An Overview of the Literature, associated Challenges, the existing Solutions, and Potential Research Directions. *arXiv preprint arXiv:2408.03335*.

Kumar, A., Gahlawat, R., Thakur, A., & Pahuja, D. (2025). A Hybrid Deep Learning Framework for IoT Network Intrusion Detection System.

Labonne, M. (2020). *Anomaly-based network intrusion detection using machine learning* (Doctoral dissertation, Institut Polytechnique de Paris).

Lindstedt, H. (2022). Methods for Network Intrusion Detection: Evaluating Rule-Based Methods and Machine Learning Models, on the CIC-IDS2017 Dataset.

Najafabadi, M. M., Khoshgoftaar, T. M., Kemp, C., Seliya, N., & Zuech, R. (2014, November). Machine Learning for detecting brute force attacks at the network level. In *2014 IEEE International Conference on Bioinformatics and Bioengineering* (pp. 379-385). IEEE.

Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, *10*, 112392-112415.

Revathy, G., Rajendran, V., Sathish Kumar, P., Vinuharini, S., & Roopa, G. N. (2022, May). Smurf attack using a hybrid machine learning technique. In *AIP Conference Proceedings* (Vol. 2463, No. 1, p. 020015). AIP Publishing LLC.

Roy, P., Kumar, R., & Rani, P. (2022, May). SQL injection attack detection by a machine learning classifier. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 394-400). IEEE.

Rustam, F., Mushtaq, M. F., Hamza, A., Farooq, M. S., Jurcut, A. D., & Ashraf, I. (2022). Denial of service attack classification using machine learning with multi-features. *Electronics*, *11*(22), 3817.

Wang, M., Yang, N., Gunasinghe, D. H., & Weng, N. (2023). On the robustness of ML-based network intrusion detection systems: An adversarial and distribution shift perspective. *Computers*, *12*(10), 209.

Zeebaree, S. R., Jacksi, K., & Zebari, R. R. (2020). Impact analysis of SYN flood DDoS attack on HAProxy and NLB cluster-based web servers. *Indones. J. Electr. Eng. Comput. Sci*, *19*(1), 510-517.

Bibliographic information of this paper for citing:

Allam, Fatima Zohra, Bousbia-Salah, Hicham, Zendaoui, Fairouz & Hamami-Mitiche, Latifa (2025). AI-Enhanced Intrusion Detection: Integrating Expert Knowledge and Machine Learning for Enterprise Networks. *Journal of Information Technology Management*, 17 (4), 99-116. https://doi.org/10.22059/jitm.2025.105485