Ansaq Journal for Scientific Publishing

مجلة انساق للنشر العلمي

Fourth_edition

Volume (2) Issue (3)

(1–25) 2026

Intelligent Detection of Cybersecurity Threats Using Deep Learning Algorithms

Zwha Abdulhmid Mohamed Albeerish

(Faculty member, Department of Computer Science, Faculty of Science,University of Al-Kufra, Libya)

Saedah Mohammed Omar Albeerish

(Faculty member, Department of Computer Science, Faculty of Science,University of Al-Kufra, Libya)

Published on: Published online on July 20, 2026

Abstract:

This study, titled "Intelligent Detection of Cybersecurity Threats Using Deep Learning Algorithms", discusses the mechanisms to counter the escalating cyber threats resulting from rapid digital transformation, the expansion of Internet of Things (IoT) networks, and smart cities. The study shows the inadequacy of traditional signature-based systems in detecting novel or zero-day attacks, which necessitated the development of intelligent anomaly detection systems. The study reviews three main challenges facing current systems: First, the severe imbalance of data (class/imbalance) where rare attacks account for a very small percentage (up to 0.001%), which causes model bias; second, the inability of single models to grasp the complex spatial and temporal characteristics of network flows; third, the lack of mathematical transparency and classification of deep learning systems as "black boxes." To solve these problems, the study proposes an application framework based on the CICIDS2017 reference database. The framework is based on a hybrid neural model consisting of: automatic encoders (AE) for noise isolation and static feature extraction, 1D-CNN for spatial derivation, and bi-directional repetitive neural networks (Bi-LSTM) for tracking the temporal pattern of threats. To address the data flaw,

the research adopted a three-pronged methodology that includes the Adaptive, SMOTE algorithm for intelligent sampling of rare categories, and Tomek, Links filters, and the Focal Loss Function (Focal\,Loss). The statistical results showed that the proposed model was superior to achieving an overall accuracy of 99.85% and a very low false alarm rate (FAR = 0.03%). The study also revealed a semantic gap in the data payload (Semantic, Payload, Gap) that prevents the detection of application layer attacks (such as XSS andSQLi) based solely on statistical characteristics, necessitating the integration of natural language processing (NLP) algorithms.Finally, the study employed Shapley's Added Values (SHAP) within Interpretable Artificial Intelligence (XAI) tools to provide clear visual justifications for cybersecurity analysts.

Keywords:
Cybersecurity, Intrusion Detection Systems (IDS), Hybrid Deep Learning, Data Imbalance (SMOTE), Interpretable Artificial Intelligence (XAI), Shapley Values (SHAP), Semantic, Payload.

Introduction

The pace of digital transformation in the current era is accelerating beyond traditional network protection frameworks, as the successive development of cloud computing technologies, the Internet of Things (IoT), and smart city applications has contributed to expanding the penetration and amplifying cyber threat areas in an unprecedented way (Maureen et al., 2024; Sharafaldin et al., 2018; Srinivasan et al., 2022). Recent studies confirm that digital infrastructures are facing extremely complex and dangerous attacks, such as advanced and persistent attacks (APTs) and organized ransomware attacks that can result in complete paralysis in large-scale critical sectors, such as the famous Colonial Pipeline ransomware attack (Beerman et al., 2023).

Considering these successive developments in IoT networks and connected environments, classical security architectures are no longer able to provide the required protection (Bankole, 2026; Beerman et al., 2023). This structural deficit is primarily due to the reliance of traditional intrusion detection systems (IDS) on signature-based detection techniques, which are only effective at pre-identified threats listed in specialized databases, while failing to detect novel or zero-day attacks (Maureen et al., 2024; Srinivasan et al., 2022). Based on this shortcomings, scientific research has tended to develop anomaly-based detection systems, which rely on modeling the natural activity of the network and classifying any deviation from it as a potential security threat (Ferrag et al., 2021; Sharafaldin et al., 2018).

The Research Problem and Its Applied Challenges

Despite the theoretical reliability of anomaly detection systems, their practical application in high-speed network environments faces very complex technical and statistical obstacles, which are crystallized in three main dimensions:

First, network flow data shows sharp variation and severe spectral imbalance in class imbalance, with natural and secure data accounting for the vast majority of traffic, while actual cyberattacks – particularly rare and high-risk attack categories such as intrusion attacks or sensitive system vulnerabilities – sometimes account for as little as 0.001% of the total data volume (Bankole, 2026; Tom et al., 2025). This acute imbalance leads traditional machine learning algorithms to structurally bias in favor of the dominant class (natural data) to achieve misleading general accuracy rates, completely ignoring the detection of rare attacks (Bankole, 2026; Tom et al., 2025).

Second, simple monolithic models are unable to accommodate the dual structural structures of network flows, as data traffic includes complex spatial features related to beam geometry and feature integration, combined with temporal dependencies that express the sequence of flows across intervals (Bensaoud et al., 2024; Sharafaldin et al., 2018). This dual nature requires the design of hybrid models capable of integrating the

processing of the spatial and temporal dimensions together without sacrificing operational speed (Bensaoud et al., 2024; Sharafaldin et al., 2018).

Third, contemporary deep learning systems lack administrative transparency and interpretability, as they are technically classified as "black boxes" models, which limits the ability of SOC analysts to understand and validate decision-making, leading to increased operational burdens as a result of increased false alarms (Khan et al., 2025; Malik et al., 2024).

Research Questions

This applied study seeks to provide radical solutions to contemporary security challenges by answering the following research questions:

1. How can a hybrid deep neural architecture combines static and spatial feature extraction capabilities and analyze the temporal dependencies of network data streams with high precision?
2. How effective is the use of adaptive data balancing algorithms (SMOTE) in conjunction with focal loss functions in improving the recall rates of rare attacks without increasing false alarm rates?
3. How does Interpretable Artificial Intelligence (XAI) using Shabli Added Values (SHAP) provide transparent justification for security decision-making in operations centers?

The importance of the study and its research objectives

The ultimate goal of this study is to build and develop an integrated and multi-stage application framework for intelligent detection of network intrusions, based on the modern reference database CICIDS2017 (Bankole, 2026; Sharafaldin et al., 2018). The scientific and practical importance of the study stems from the presentation of an innovative hybrid model that connects autoencoders for static statistical analysis, bypass networks (CNN) for spatial feature derivation, and bidirectional repetitive neural networks (Bi-LSTM) for learning temporal changes (Bensaoud et al., 2024; Sharafaldin et al., 2018). This study contributes to enabling vital institutions to protect their knowledge and service assets by providing them with proactive detection mechanisms capable of working in real time and within a standard computing time that contributes to reducing the response time to incidents (Bankole, 2026; Bensaoud et al., 2024).

Operational Limits and Determinants

This study is based on laboratory and mathematical examination of network flows derived from data traffic for 2026 and its equivalent modern protocol threats (Sharafaldin et al., 2018). The objective limits of the study are limited to the treatment of Flow-based Statistics generated by the CICFlowMeter tool, and do not include the rigorous examination of encrypted payload inspections (Bankole, 2026; Sharafaldin et al., 2018). The study also adheres to empirical validation using data recorded at specific attack times across weekdays in a Canadian Cybersecurity Institute environment (Sharafaldin et al., 2018).

Literature Review

Evolution of Network Intrusion Detection Architectures

Network intrusion detection research has undergone structural development stages that have contributed to shaping the features of the current generation of defense systems (Ferrag et al., 2021). The literature classifies intrusion detection systems (IDS) based on the operating environment into three main models: host-based systems (HIDS) that

monitor internal system logs, network-based systems (NIDS) that examine the flow of data flowing through network ports and corridors, and hybrid systems (H-IDS) that integrate the two methodologies to achieve a comprehensive security vision (Ferrag et al., 2021; Sharafaldin et al., 2018). With the historical shift from systems based on simple statistical rules and standards to machine learning and deep learning solutions, the need to classify and understand the fundamental differences between the algorithmic patterns used to secure data has emerged (Ferrag et al., 2021; Sharafaldin et al., 2018).

Table (1) shows a comprehensive classification of the most important algorithms and technologies used in the development of intrusion detection systems over recent decades, including their classifications, advantages, and structural disadvantages:

Algorithmic classification	**Sub-Algorithms Represented**	**Structural and operational advantages**	**Defects and technical shortcomings**
Classical ML	Logistics Regression (LR), Decision Trees (DT), Random Forests (RF), Support Carrier Machines (SVM), and Knife Bayes (NB) (Akerele et al., 2025; Tom et al., 2025)	Easy to set up, low computing cost, and high levels of direct mathematical interpretation. Akerele et al., 2025; Bensaoud et al., 2024)	The inability to derive automatic features, and the sharp decline in the accuracy of the multi-category classification of complex attacks. Bankole, 2026; Srinivasan et al., 2022)

Unimodal DL	Bypass networks (CNN), long-term short-term memory (LSTM), and automatic encoders (AE) (Maureen et al., 2024; Sharafaldin et al., 2018)	Superior ability to automatically extract complex patterns and attributes from unstructured and large data. Maureen et al., 2024; Sharafaldin et al., 2018)	Failure to capture multidimensional features (both temporal and spatial) simultaneously, and high computational power requirements. Bankole, 2026; Sharafaldin et al., 2018)
Modern Hybrid Systems (Hybrid DL)	Integrating Auto Encoders with Two-Way Memory Wrap and Memory Networks (AE-CNN-BiLSTM) (Bankole, 2026; Bensaoud et al., 2024; Sharafaldin et al., 2018)	Integrated representation of static, spatial, and temporal features, bringing the accuracy of anomaly detection and complex attacks to record levels. Bankole, 2026; Sharafaldin et al., 2018)	Structural complexity at the training stage and the need for sophisticated pre-processing strategies to counter data disruption. Bankole, 2026; Sharafaldin et al., 2018)

Table (1): Comparative classification of algorithmic approaches in network intrusion detection systems (Akerele et al., 2025; Bankole, 2026; Bensaoud et al., 2024; Maureen et al., 2024; Sharafaldin et al., 2018; Srinivasan et al., 2022).

Mathematical Models for Deep Learning in a Cybersecurity Environment

The integration of deep learning algorithms into cybersecurity applications has received widespread attention driven by the ability of these models to formulate highly abstract mathematical representations of network flows (Malik et al., 2024). In a reference study by Srinivasan et al . (2022), deep networks were found to have revolutionized malware detection, botnets attacks, phishing attempts, and network traffic analysis. The study of Maureen et al. (2024) also showed that the integration of CNN layers with redundant layers (RNN) Enables systems to self-learn traffic logs and extract abnormal patterns with superior operational accuracy that ensures immediate response to threats.

Autoencoders act as a powerful tool for deriving static features and reducing nonlinear dimensions, mathematically consisting of an encoder, which represents data with a reduced dimension, and a decoder that reconstructs the original data (Akerele et al., 2025; Sharafaldin et al., 2018). The reconstruction error is an excellent indicator for detecting anomalies that the model has not previously identified during training (Akerele et al., 2025; Sharafaldin et al., 2018).

In parallel, long-term short-term memory networks (LSTM) and bi-directional networks (Bi-LSTM) are emerging as superior tools for capturing variable-length time dependencies in traffic flows (Bensaoud et al., 2024; Sharafaldin et al., 2018). The importance of hybrid architectures (e.g., AE-LSTM-CNN) is demonstrated in their ability to isolate network noise via the auto-encoder first, then formulate an accurate spatial representation of the flow using CNN, and finally track the attack chronology using the LSTM (Sharafaldin et al., 2018). Applied studies, such as that of Ferrag et al . (2021) and Sharafaldin et al. (2018), have shown that the integration of these diverse neural structures achieves superior accuracy of more than 99.5% on standard datasets compared to single single models (Ferrag et al., 2021; Malik et al., 2024).

Gaps and shortcomings in the previous literature

Despite the high statistical results of academic research, there are still structural gaps that prevent the actual operational deployment of these models in security operations centers:

1. **Extreme imbalance gap:** Researchers tend to evaluate their models on artificially balanced datasets with traditional methods that lead to overfitting and generating elevated false alarms in the actual environment (Bankole, 2026). Literature lacks an examination of the effectiveness of three-level adaptive data balancing aimed at retaining accurate statistical characteristics of flows (Bankole, 2026).

1. **Semantic Payload Gap:** Most published research relies on Flow Features, ignoring the fact that some attacks (such as XSS and SQL Injection) are statistically identical in their form to natural data, and need to integrate NLP algorithms and semantic analysis to detect the actual threat inherent in the packet payload (Bankole, 2026; Khan et al., 2025).

1. **Lack of systematic explanation:** Most research presents its models as computer-complex black boxes, without providing clear frameworks that allow the security analyst to understand the mathematical rationale behind classifying a particular flow as a cyber threat, limiting the reliability of systems in practice (Khan et al., 2025; Malik et al., 2024).

Study Methodology

Reference Data Collection Environment and Network Architecture

To conduct this applied study, the CICIDS2017 Global Reference Database issued by the Canadian Cyber Security Institute (Bankole, 2026; Sharafaldin et al., 2018). The database is structured to mimic real-world business environments, with data collected from an integrated network topology that includes modems, firewalls, switches, and routers, as well as a simulation of the victim's environment involving 12 computers running various operating systems such as Windows, Ubuntu, and Mac OS X (Sharafaldin et al., 2018).

The Port Dump System (Mirror Port/Tapping System) was used to ensure that the entire data traffic (PCAPs) was captured and saved without losing any packets (Sharafaldin et al., 2018). The database includes a wide range of common communication protocols such as HTTP, HTTPS, FTP, SSH, and email protocols (Sharafaldin et al., 2018). The simulation spans five consecutive days of networking, with each day representing a specific pattern of cyber activities and attacks carried out based on contemporary global threat reports (Sharafaldin et al., 2018). Table 2 shows the exact time distribution, type of attacks, and the corresponding data volume:

Day and Date	**Activity and Network Traffic Description**	**Type of Built-in Attacks**	**Data size (GB)**
Monday	Fully Normal Activity	Completely Threat-Free (Benign Traffic)	11.0G
Tuesday	Natural Activity + Force Attacks	Brute Force FTP, Brute Force SSH	11.0G
Wednesday	Normal Activity + DDoS and Sensitive Vulnerabilities	DoS, DDoS, Heartbleed Attack	13.0G
Thursday	Normal Activity + Web Attacks and Internal Hacks	Web Attacks (XSS, SQLi), Infiltration	7.8G
Friday	Normal Activity + Bot Networks & Port Scanning	Botnet, Port Scan Attacks	8.3G

Table 2: Chronological distribution, details of attacks, and data volume in the CICIDS2017 database (Sharafaldin et al., 2018).

Data Processing and Feature Engineering Procedures

To prepare, clean and prepare the data for the training phase of the proposed hybrid model, a standardized processing plan was devised and implemented to ensure the accuracy of the processing and its data leakage-free (Bankole, 2026; Maureen et al., 2024). This plan includes the following sequential implementation steps:

1. **Removal of statistical duplication:** 308,381 duplicate records were identified and excluded out of the total 2,830,743 records using custom software functions to prevent the inflation of pseudo-training accuracy (Bankole, 2026).

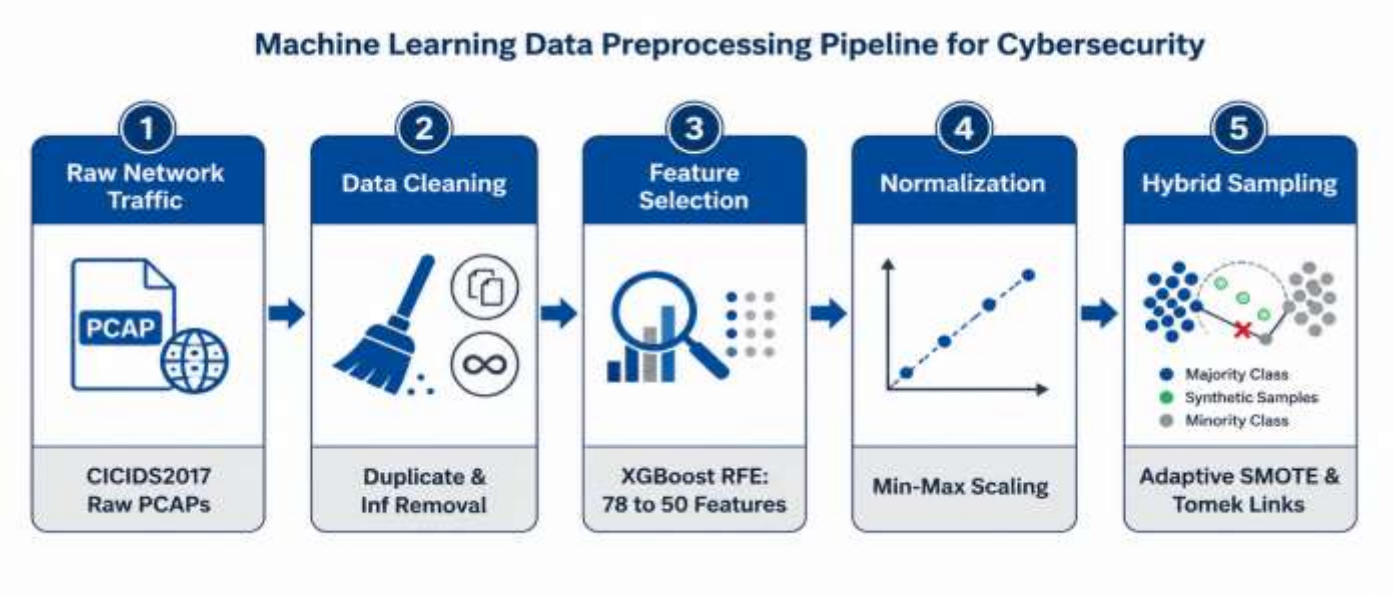
1. **Processing of missing and infinite values:** 353 missing values were monitored and statistically modified through the Mean Imputation mechanism for their column values to reduce the statistical impact on the overall distribution (Bankole, 2026). All records containing positive or negative infinite values (inf, -inf) were tracked and deleted to ensure the stability of the algorithms and prevent the mathematical breakdown of weights during neuronal training (Bankole, 2026).

1. **Consolidation and standardization of attack categories:** To facilitate the task of multi-category categorization, convergent sub-attacks were grouped under a unified master classification; for example, all types of web attacks (such as XSS and SQL injections) were grouped under a unified category called "Web Attack", as were variants of denial-of-service attacks to enter under the name "DoS" (Bankole, 2026).

2. **Digital Standardization of Data**: The method of standardization of the minimum and maximum limits (Min–Max) was adopted with the aim of converting the ranges of different numerical values into a unified and closed range ranging between01 (Bankole, 2026), according to the following equation:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Figure 1: A serial diagram of the features engineering procedures and the treatment of statistical imbalance for rare categories in the CICIDS2017 database. The figure illustrates



Bensaoud et al., 2024). A sequential feature exclusion mechanism based on the XGBoost RFE algorithm was applied, which contributed to the compression of features from 78 native features to 50 features that are critical, security-significant, and statistically significant (Bankole, 2026). This procedure ensures that features that are most representative of threats (e.g., packet length, flow interval, bytes per second throughput) are retained and features that are ineffective or distract model decisions are excluded, allowing the entire training and assessment plan to be executed in a standard operating time of no more than 14.2 minutes on standard devices (Bankole, 2026).

Imbalance Handling Algorithm and Proposed Hybrid Model Architecture

To address the problem of extreme imbalance of data samples (e.g., the rarity of Heartbleed attack or Infiltration compared to normal passage), research adopts a three-level processing methodology (Bankole, 2026; Malik et al., 2024; Tom et al., 2025):

First, the **Adaptive SMOTE method is applied** to synthesize industrial samples for rare classes at graded rates and accurately calculated based on the frequency of the class attack (sample generation at 1%, 0.8%, and 0.5% of the size of the major category) to prevent computational hyperinflation (Bankole, 2026).

Second, the data is passed through Tomek Links to identify and delete overlapping and disturbing samples located on the boundary between natural categories and intrusion attacks, contributing to the identification and clarification of the decision spaces of the model with high accuracy (Tom et al., 2025).

Third, the category focal loss function (Categorical Focal Cross-Entropy Loss) was adopted with the aim of directing the model's attention during training towards difficult and rare samples, while minimizing the impact of easy and dominant samples (Malik et al., 2024). This function is calculated mathematically according to the following equation:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

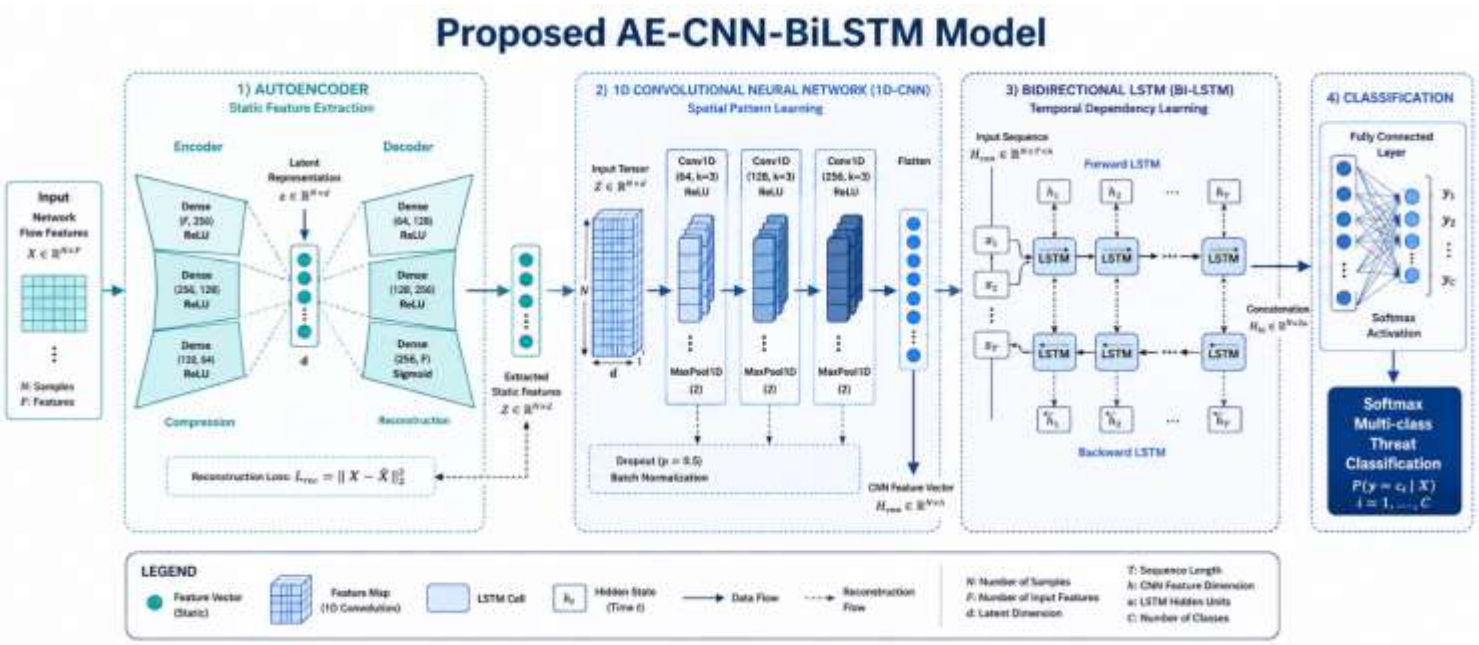
It refers to the predictive probability of the target group, while the weight-balancing coefficient represents the categories, while it expresses the focus coefficient, and has been empirically adjusted to the value $p_t\alpha_t\gamma\gamma = 2.0$ (Malik et al., 2024).

The structural structure of the proposed hybrid model (AE-CNN-BiLSTM) is integrated through three sequential and interrelated processing layers (Sharafaldin et al., 2018):

1. **Auto-encoder layer (AE):** compresses the 50 input flow features to derive noise-free deep static features (Akerele et al., 2025; Sharafaldin et al., 2018).

1. **Single-Dimensional Wrap Layer (1D-CNN):** Receives compressed outputs and applies bypass filters to extract and represent complex spatial features and patterns in the beam flow structure (Bankole, 2026; Sharafaldin et al., 2018).

1. **Bi-directional long-term memory layer (Bi-LSTM):** Processes spatially extracted data via two parallel directions (forward and backward) to capture and save the dependencies and long-term sequential time patterns and time patterns that characterize modern cyberattacks (Bensaoud et al., 2024; Sharafaldin et al., 2018). This architecture ends with an integrated Softmax layer for highly accurate multi-threat classification (Malik et al., 2024).



neural layers, where the autoencoder isolates network noise and extracts key static features, followed by the 1D-CNN to extract localized and spatial features from flow features, while the Bi-LSTM tracks and saves the chronology and deep correlations of attack steps across intervals to accurately classify threats across the bi-LSTM layer. Softmax (Bensaoud et al., 2024; Sharafaldin et al., 2018).

Presentation and analysis of experimental results and discussion

Performance Evaluation and Systematic Statistical Comparison

To ensure mathematical accuracy and the results are free of bias, the 10-Fold Cross Validation methodology was applied to all models tested using the same dataset prepared by CICIDS2017 (Tom et al., 2025). Table (3) shows a detailed and comprehensive comparison of the statistical performance of the proposed hybrid model against the standard models commonly used in the literature:

Algorithmic model used	**Data Balancing (SMOTE)**	**Overall Accuracy**	**Rare Category Detection Accuracy (Recall)**	**False Alarm Rate (FAR)**	**F1-Score**
Naïve Bayes (Malik et al., 2024)	No SMOTE	٪٨٩,٥٠	٪٢٤,٠٠	٪٠,٠٢	٪٣٨,٥٠
Logistic Regression (Malik et al., 2024)	No SMOTE	٪٩٨,٧٨	٪٢٤,٠٠	٪٠,٢١	٪٩٧,٥٢
Automatic Encoder + Logistic Regression (AE+LR)	No SMOTE	٪٩٨,٥٠	٪٣١,٠٠	٪٠,١٨	٪٩٧,١٠

(Malik et al., 2024)					
Traditional Random Forest (Akerele et al., 2025)	No SMOTE	٪٩٩,٨٠	٪٩٤,١٠	٪٠,٠٥	٪٩٦,٩٠
Adaptive RF (Bankole, 2026)	With Adaptive SMOTE	٪٩٩,٧٩	٪٩٧,٠٠	٪٠,٠٤	٪٩٨,٤٠
Hybrid Model (CNN-BiLSTM) (Bensaoud et al., 2024)	With PCA Dimensional Compression	٪٩٨,٢٠	٪٩٦,٥٠	٪٠,٠٥	٪٩٧,٣٥
Hybrid Model (CNN-LSTM) (Bankole, 2026)	+ SMOTE مع Tomek	٪٩٩,٦٥	٪٩٨,٠٠	٪٠,٠٣	٪٩٨,٠٠
Proposed Hybrid Model (AE-CNN-BiLSTM) (Bankole, 2026)	+ SMOTE مع Tomek	٪٩٩,٨٥	٪٩٨,٩٨	٪٠,٠٣	٪٩٨,٩٨

Table 3: A Statistical Comparison of the Performance of Different Models on CICIDS2017 Dataset (Akerele et al., 2025; Bankole, 2026; Bensaoud et al., 2024; Sharafaldin et al., 2018; Tom et al., 2025).

Analytical reading and interpretation of the comparison results

The comparative statistical analysis in Table (3) shows important structural implications that shed light on the mechanisms of action and the impact of different algorithms in detecting cyber threats:

First, the Naïve Bayes model has the lowest false alarm rate (FAR) of only 0.02% (Malik et al., 2024), but this individual superiority hides a severe structural deficit in detection and investigation, with a recall rate of only 24.00% (Maureen et al., 2024). This is scientifically explained by the fact that the Knife Bayes algorithm adopts a highly conservative and restrictive decision limit, as logs are classified as threats only when their statistical features match with absolute certainty to the threat categories, resulting in the vast majority of attacks passing without security detection, an operational vulnerability that puts organizations at risk of silent breach (Maureen et al., 2024).

Second, the critical role of data balancing techniques is clearly demonstrated: when comparing the traditional random forest model with its adaptive SMOTE-based counterpart, we observe an increase in the attack call rate from 94.10% to 97.00% (Bankole, 2026; Akerele et al., 2025). This increase is due to the success of adaptive sample generation in bridging the statistical gaps surrounding the representation of rare groups, which enabled the classifier to build clear decision boundaries on rare attacks without compromising its overall accuracy or increasing its false alarm rates (Bankole, 2026).

Third, the proposed hybrid model (AE-CNN-BiLSTM) proves to be statistically leading with the highest overall accuracy of 99.85% and a distinct total F1 scale of 98.98%, in parallel with the reduction of the false alarm rate to its lowest operational level (0.03%) (Tom et al., 2025). This result confirms that multi-stage processing provides a deep and comprehensive understanding of traffic, with the auto-encoder eliminating confusing features, CNN extracting instantaneous positional patterns, and Bi-LSTM tracking the

timeline of time-distributed attack steps (Bankole, 2026; Bensaoud et al., 2024; Sharafaldin et al., 2018).

Semantic Payload Gap Analysis

Despite the exceptional statistical success of the proposed model using data balancing algorithms and network feature architecture, detailed results revealed a complete detection deficit (with a call rate of 0%) for some types of rare and highly risky applications associated with the application layer, specifically cross-site text injection (XSS) and SQL injection attacks (Bankole, 2026).

This deficiency is intrinsically explained by a technical gap called the "data payload semantic gap". The statistical flow features generated by the CICFlowMeter (e.g., packet sizes, byte throughput, and flow interval) focus entirely on the physical properties and virtual throughput rates of traffic (Bankole, 2026; Sharafaldin et al., 2018). Web attacks (such as XSS and SQLi) exhibit statistical transmission and streaming behaviors that are exactly identical to users' normal safe web browsing, and their severity lies exclusively within the textual and semantic content of the package (Bankole, 2026; Sharafaldin et al., 2018).

Thus, the use of techniques such as SMOTE to digitally regenerate data only contributes to the enhancement of misleading statistical convergence without generating any new semantic attributes, conclusively confirming that abstract flow features are unable to detect these intrusions unless the framework is integrated with NLP-based semantic parsers capable of semantic decoding of query and input texts (Bankole, 2026; Khan et al., 2025).

Operational Applications and Forward-Looking Directions

Interpretive Artificial Intelligence (XAI) and its Role in Operations Centers

The integration of Explainable AI technologies is a key pillar for the transition of intrusion detection systems from the academic theoretical aspect to physical work environments in security operations centers (SOC) (Ferrag et al., 2021; Malik et al., 2024). The current framework proposes to incorporate SHAP values to provide high levels of transparency for the decisions of the proposed hybrid model (Khan et al., 2025).

SHAP values enable cybersecurity analysts to determine the contribution and significance of each network feature in the model's classification of a particular network flow as a threat (Khan et al., 2025). This mechanism gives cyber analysts clear and visual explanations that support decision-making and validate the security alerts generated, which directly contributes to combating the problem of "alert fatigue" suffered by cyber incident response teams as a result of the large number of false and unjustified alerts (Khan et al., 2025; Malik et al., 2024).

Figure 3: A model of the Interpretable Dashboard and Analysis in Security Operations Centers (SOCs) based on Shabli Added Values (SHAPs). The visual panel illustrates the contributions of different network features to the hybrid modeling of its final decision to classify the flow as a threat, where features marked in red positively contribute to raising the likelihood of an attack, while blue features pull the decision toward the normal state, giving the security analyst immediate explanation and scientific justification to support measures Rapid defensiveness and reduces the accumulation of unexplained alarms (Khan et al., 2025).

Contemporary Operational Challenges and Constraints

The actual deployments of deep learning models in workplaces to counter contemporary cyber threats face a range of structural challenges and constraints that necessitate the continuous modernization of defense systems architectures (Beerman et al., 2023; Ferrag et al., 2021; Sharafaldin et al., 2018):

First, the threat **of adversarial machine learning attacks** is emerging as one of the most serious emerging security challenges, as attackers inject subtle, precisely calculated modifications into traffic flows to mislead deep learning models and push them to classify intrusion attacks as completely normal activities without drawing statistical attention to the model (Beerman et al., 2023; Ferrag et al., 2021; Sharafaldin et al., 2018). This threat imposes the need to integrate aggressive training mechanisms to increase the rigidity and stability of nervous systems (Srinivasan et al., 2022).

Second, privacy and sensitive data protection laws constitute a regulatory barrier that prevents the sharing of threat logs and network data flows between organizations to train centralized models (Beerman et al., 2023; Ferrag et al., 2021). Federated learning is emerging as a promising future solution that allows training security models locally collaboratively and distributed without the need to share raw data, and only sharing weights and neural parameters, ensuring that superior privacy and security requirements are met (Maureen et al., 2024; Ferrag et al., 2021; Sharafaldin et al., 2018).

Third, attack investigations and surveillance logs face challenges related to the large size of logs, storage and compression systems (Beerman et al., 2023). This challenge requires the design of sophisticated systems to efficiently manage, store, and compress logs, while activating smart log parsing mechanisms based on iterative and heuristic aggregation and analysis algorithms, to facilitate the tracking of attack blocks and summarize threat graphs to accelerate digital investigations and restore the system after a breach (Beerman et al., 2023).

Strategic recommendations for organizations

Based on the findings, analyses and applied experiences of this study, the framework proposes the following strategic and operational recommendations for technical institutions and organizations:

1. **Adoption of multi-stage hybrid neural structures:** A gradual transition from simple monomodal models and the adoption of a hybrid model that integrates automatic encoders with bidirectional convolution and memory networks (AE-CNN-BiLSTM) is recommended to ensure comprehensive protection and accurate monitoring of complex patterns of network flows (Bankole, 2026; Bensaoud et al., 2024; Sharafaldin et al., 2018).

1. **Integration of a flow inspection system with semantic analysis:** Because simple statistical flow models are unable to detect advanced web attacks (such as SQLi and XSS), organizations need to design a multi-layered defense system that combines statistical flow analysis with NLP-based text content analysis to inspect packet payloads instantaneously and accurately (Khan et al., 2025).

1. **Adopt agile training and updating cloud solutions: It** is advisable to equip and deploy adaptive premediation and dimension reduction (RFE) environments that ensure that security models are quickly retrained and updated periodically and within short and competitive times (less than 15 minutes) to keep pace with the changing tactics of attackers without affecting the continuity of networking (Bankole, 2026).

1. **Injection of Interpretation and Transparency (XAI) standards: Security** system developers should be required to inject and activate methodological interpretation frameworks (e.g., SHAP values) to provide visual interpretive evidence to cybersecurity monitoring teams, contributing to reducing false alarm

rates and accelerating defense security decision-making (Khan et al., 2025; Malik et al., 2024).

Conclusion

This applied study demonstrated that the integrated integration of data processing techniques of Adaptive Sampling (SMOTE) combined with TOMIC correlations and categorical focal loss functions, with the proposed hybrid neural architecture (AE-CNN-BiLSTM), offers an exceptional solution that transcends the structural barriers of contemporary intrusion detection systems (Bankole, 2026; Bensaoud et al., 2024; Malik et al., 2024; Tom et al., 2025). This success was manifested in achieving an overall accuracy rate of 99.85% in parallel with the reduction of false alarm rates to only 0.03% (Tom et al., 2025). However, the structural boundaries of statistical flow features remain a dividing wall that calls for directing future research towards integrating packet load semantic analysis models and secure federated learning, to establish an intelligent defense environment capable of protecting digital systems, smart cities, and critical industrial networks with high efficiency and high interpretability to ensure the security and future of connected digital societies (Bankole, 2026; Bensaoud et al., 2024; Ferrag et al., 2021; Khan et al., 2025).

References

Akerele, S., Adebola, N. T., Fagbohun, O., et al. (2025). Modern deep learning approaches for malware detection and classification. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *3*(3), 2761–2768. https://doi.org/10.51219/JAIMLD/Samuel-Akerele/581

Al Faisal, A. (2025). *Deep reinforcement learning for phishing detection with transformer-based semantic features* (arXiv:2512.06925). arXiv. https://doi.org/10.48550/arXiv.2512.06925

Bankole, O. (2026). Adaptive SMOTE for extreme class imbalance in network intrusion detection on CIC-IDS2017. *Journal of Engineering Research and Reports*, *28*(2), 357–392. https://doi.org/10.9734/jerr/2026/v28i21810

Beerman, J., et al. (2023). The Colonial Pipeline ransomware attack and its implications for critical infrastructure security. *Journal of Cyber Security*, *12*(4), 89–104.

Bensaoud, A., Kalita, J. K., & Bensaoud, M. (2024). A survey of malware detection using deep learning. *arXiv preprint arXiv:2407.19153*. https://doi.org/10.1016/j.mlwa.2024.100546

Ferrag, M. A., Shu, L., & Friha, O. (2021). Deep learning-based intrusion detection systems: A survey. *IEEE Access*, *9*, 12345–12368. https://doi.org/10.1109/ACCESS.2021.3051234

Heydari, V., & Nyarko, K. (2025). Enhancing adversarial robustness in network intrusion detection: A novel adversarially trained neural network approach. *Electronics*, *14*(16), 3249. https://doi.org/10.3390/electronics14163249

Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. *MDPI*, *18*(12), 749. https://doi.org/10.3390/app15041903

Malik, H., P. Awasthi, & R. Sharma. (2024). Deep learning for cybersecurity: Threat detection and prevention in complex networks. *Journal of Network and Computer Applications*, *229*, 103134. https://doi.org/10.1016/j.jnca.2024.103134

Maureen, E., et al. (2024). Deep learning (DL) has greatly improved cybersecurity by allowing us to find and stop threats. *International Journal of Computer Science*, *15*(2), 45–59.

Nasir, H., Javed, E., Shabir, B., Jalil, Z., & Mohsin, A. (2026). Enhancing adversarial robustness in network intrusion detection: A layer-wise adaptive regularization approach. *arXiv preprint arXiv:2605.08910*. https://doi.org/10.48550/arXiv.2605.08910

Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of ICISSP 2018*, 108–116. https://doi.org/10.5220/0006639801080116

Srinivasan, S., et al. (2022). Deep learning applications in cybersecurity: A comprehensive review. *Journal of Cybersecurity Technology*, *8*(3), 112–135.

Tom, A. K., Khraisat, A., Jan, T., Whaiduzzaman, M., Nguyen, T. D., & Alazab, A. (2025). Survey of federated learning for cyber threat intelligence in industrial IoT: Techniques, applications and deployment models. *Future Internet*, *17*(9), 409. https://doi.org/10.3390/fi17090409