A Hybrid CTGAN-SMOTE and VAE-LSTM Framework for Interpretable Intrusion Detection in Imbalanced Network Traffic

Felicia Maake , Justice Nkoana * , Vekani Reviet Baloyi * , Sello Mokwena *

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1813.v1

Keywords: GenAI; intrusion detection; class imbalance; CTGAN; SMOTE; LSTM; VAE; SHAP; cybersecurity

 doi:10.20944/preprints202605.1813.v1

Article

A Hybrid CTGAN-SMOTE and VAE-LSTM Framework for Interpretable Intrusion Detection in Imbalanced Network Traffic

Felicia Maake [1], Justice Nkoana [2,*], Vekani Reviet Baloyi [1] and Sello Mokwena [1,*]

[1] Department of Computer Science, University of Limpopo, Polokwane, South Africa

[2] Department of Information Systems, University of Cape Town, Rondebosch, South Africa

* Correspondence: nknphu010@myuct.ac.za (J.N.); sello.mokwena@ul.ac.za (S.M.)

Abstract

The increasing sophistication of cyber threats poses significant challenges to traditional intrusion detection systems, particularly in the presence of highly imbalanced network traffic. This study aims to develop a hybrid intrusion detection framework that improves detection performance while maintaining model interpretability. The proposed approach integrates data augmentation, deep learning and explainable artificial intelligence within a unified pipeline. Specifically, Synthetic Minority Over-Sampling Technique (SMOTE) and Conditional Tabular Generative Adversarial Networks (CTGAN) are employed to generate realistic samples for minority attack classes. A Long Short-Term Memory (LSTM) network is used to capture temporal patterns in network traffic, while a Variational Autoencoder (VAE) provides probabilistic anomaly validation. The model is evaluated on the CICIDS 2018 dataset, achieving an accuracy of 99.08% and a ROC-AUC score of 0.9949. To enhance transparency, SHapley Additive exPlanations (SHAP) are applied, identifying source and destination ports and TCP flags as key contributing features. This explicit feature attribution proves that the model relies on legitimate network indicators rather than synthetic noise or dataset artifacts. The results indicate that the proposed hybrid framework effectively addresses class imbalance and improves detection performance while providing interpretable insights suitable for operational cybersecurity environments.

Keywords: GenAI; intrusion detection; class imbalance; CTGAN; SMOTE; LSTM; VAE; SHAP; cybersecurity

1. Introduction

The global cybersecurity landscape is increasingly characterized by the rapid evolution of cyber-attacks, which now match the pace of technological advancement [1]. Traditional security systems that rely on static rules and known signatures are frequently bypassed by emerging and sophisticated attack techniques, rendering traditional firewalls and signature-based detection insufficient [2]. In response, Generative Artificial Intelligence (GenAI) techniques have emerged as a promising approach for enhancing anomaly detection and addressing these critical vulnerabilities [3,4].

However, the practical implementation of artificial intelligence in network security is hindered by several primary challenges, most notably severe class imbalance, the inability to detect unseen threats, and a lack of model interpretability [5]. Real-world cybersecurity datasets are inherently imbalanced, with normal traffic heavily dominating while malicious events, such as advanced persistent threats (APTs) and zero-day exploits, are critically scarce [6,7]. Training models on such imbalanced data introduces severe bias toward the majority class, leaving the system blind to low-frequency but high-impact attacks [8]. Furthermore, while sequential deep learning models like Long Short-Term Memory (LSTM) networks are highly effective at capturing temporal patterns, they rely on predefined decision boundaries that struggle against novel anomalies. Compounding this issue,

 doi:10.20944/preprints202605.1813.v1

these models operate largely as black-box systems [9]. This opacity limits their interpretability and reduces trust in critical decision-making contexts where security analysts must confidently verify alerts [10,11].

To address these limitations, this study proposes a comprehensive, multi-stage hybrid framework that simultaneously tackles data imbalance, temporal dependency, probabilistic anomaly detection, and operational transparency. By integrating the Synthetic Minority Over-sampling Technique (SMOTE) and Conditional Tabular GAN (CTGAN) for robust tabular synthesis, a hybrid VAE-LSTM architecture for comprehensive threat detection, and SHAP for explainability, this research bridges critical gaps in the current literature regarding unified and balanced hybrid architectures. Specifically, this study tests the hypothesis that combining generative data augmentation (to resolve majority class bias) with probabilistic anomaly detection (to catch unseen threats) can maintain high detection precision while delivering verifiable, explainable alerts. The principal findings show that the proposed model achieves a high accuracy of 99.08% on the CICIDS 2018 dataset, and that the inclusion of SHAP values provides actionable insights into the feature-level logic of the detection engine, identifying critical protocol flags and port activities as primary indicators of intrusion.

2. Related Work

Intrusion Detection Systems (IDS) have advanced significantly through the integration of machine learning and deep learning techniques. However, critical challenges such as severe class imbalance, complex spatiotemporal feature representation, and limited model interpretability continue to hinder their effectiveness in real-world environments. Recent surveys emphasize that modern IDS must achieve not only high detection accuracy but also adaptability to evolving attack patterns and heterogeneous network conditions [12]. Historically, traditional machine learning approaches, such as Support Vector Machines (SVM) and Random Forests, provided strong baselines. However, these methods rely on static feature representations and often fail to capture the dynamic, sequential nature of modern threats, justifying the shift toward deep learning architectures [12]. Although deep learning-based frameworks have become the dominant approach due to their ability to automatically learn hierarchical representations from network traffic data, they still exhibit limitations in temporal feature learning, scalability, and robustness under imbalanced datasets [12].

To address these limitations, hybrid architectures such as CNN–LSTM models have been extensively investigated. In more recent frameworks, CNNs extract spatial correlations in network flows, while LSTM and GRU networks capture sequential dependencies over time. Enhanced variants incorporating attention mechanisms and CNN–LSTM–GRU combinations have demonstrated improved detection performance in dynamic and complex network environments [13–15]. Nevertheless, these models remain highly sensitive to skewed class distributions, resulting in degraded performance on rare attack classes.

While recent literature has increasingly adopted Transformer-based models for network anomaly detection due to their superior capability in capturing long-range global dependencies, standard Transformer architectures inherently introduce significant computational overhead. The complexity of self-attention mechanisms often makes them computationally intensive, presenting scalability challenges that can make them unsuitable for real-time, resource-constrained intrusion detection environments [16]. Consequently, this study explicitly selects the LSTM architecture, which provides a computationally efficient baseline for processing sequential temporal dependencies while still maintaining robust anomaly detection capabilities.

To mitigate class imbalance, data-level techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have traditionally been employed. However, SMOTE relies on linear interpolation between samples, which limits its ability to represent complex nonlinear attack distributions. Consequently, generative approaches have gained increasing attention. In particular, Conditional Tabular GAN (CTGAN) has been introduced to better handle structured cybersecurity data by mitigating instability and mode collapse in conventional GAN architectures. Empirical

 doi:10.20944/preprints202605.1813.v1

studies demonstrate that CTGAN-based pipelines improve minority class representation and reduce overfitting in intrusion detection tasks [17,18].

Beyond class imbalance, discriminative learning models are inherently limited in detecting previously unseen attacks due to reliance on fixed decision boundaries. This limitation has motivated the adoption of probabilistic generative models such as Variational Autoencoders (VAEs), which learn normal behaviour distributions and identify anomalies via reconstruction error. Recent approaches integrating VAEs with distribution alignment techniques (e.g., MMD-based learning) and open-set recognition strategies demonstrate strong capability in detecting unknown attack patterns in dynamic environments [19,20].

Another major challenge in modern IDS is the lack of interpretability [21]. Deep learning-based IDS models are typically black-box systems, limiting their deployment in security-critical environments where transparency and trust are essential [22]. Explainable Artificial Intelligence (XAI) techniques have therefore become increasingly important in cybersecurity research [23]. Systematic studies highlight that integrating XAI is essential for developing trustworthy and resilient intrusion detection systems, particularly in adversarial and industrial contexts [24]. However, most existing approaches apply explainability post-hoc and rarely integrate it within unified end-to-end IDS architectures [25].

Overall, recent research trends indicate a clear transition toward hybrid and multi-stage IDS frameworks that integrate deep learning, generative modelling, and explainability [26]. However, existing studies continue to address these challenges in isolation. For instance, CTGAN-based methods enhance synthetic data generation but lack interpretability mechanisms [17,18]. Conversely, XAI-based approaches improve transparency but are typically evaluated on simpler or non-generative architectures [24]. Similarly, VAE-based anomaly detection improves probabilistic modelling for unseen attacks but is rarely combined with robust data augmentation strategies to address severe class imbalance [19,20]. While these approaches contribute meaningfully to individual components of IDS design, they primarily optimize isolated stages rather than delivering a unified framework.

While the proposed framework utilizes SMOTE and CTGAN, various alternative techniques exist in the literature. Traditional oversampling algorithms like ADASYN have been proposed to manage class imbalance; however, they frequently introduce noise when applied to highly non-linear, high-dimensional cybersecurity data [27]. Furthermore, while ensemble machine learning methods such as Random Forest, XGBoost, and Isolation Forests offer strong baseline performances for anomaly detection, they lack the intrinsic architectural capacity to extract complex spatiotemporal dependencies from sequential network flows [28]. Consequently, this study explicitly selects a hybrid VAE-LSTM architecture. The LSTM is uniquely suited to capture the deep temporal mechanics of network traffic [29], while the VAE provides a probabilistic reconstruction advantage that traditional tree-based models lack [30]. By coupling these discriminative and generative deep learning capabilities with the robust tabular synthesis of CTGAN, the chosen methodology overcomes the specific limitations of standard machine learning baselines.

Therefore, the development of a unified IDS framework that simultaneously addresses class imbalance, temporal dependency modeling, probabilistic anomaly detection, and interpretability remains an open problem in the literature. This gap motivates the proposed method, which combines SMOTE and CTGAN for robust data balancing, a hybrid VAE–LSTM architecture for enhanced detection, and XAI techniques for explainable decision-making, thereby providing a unified and interpretable solution to modern intrusion detection challenges.

3. Methodology

This section outlines the systematic methodology employed to develop, train, and evaluate the proposed hybrid generative and discriminative framework for network intrusion detection. The experimental approach is structured into a sequential pipeline comprising rigorous data preprocessing, a dual-stage SMOTE-CTGAN balancing strategy to mitigate severe class imbalance,

 doi:10.20944/preprints202605.1813.v1

and the deployment of a coupled VAE-LSTM architecture for robust anomaly identification. Finally, the framework incorporates SHapley Additive exPlanations (SHAP) to ensure model interpretability, thereby bridging the gap between high-accuracy threat detection and operational transparency.

3.1. Data Preparation and Preprocessing

This study utilized the CICIDS 2018 dataset, obtained from the Canadian Institute for Cybersecurity, which provides a realistic representation of modern network traffic, including multiple categories of cyberattacks. The dataset was preprocessed to ensure consistency and suitability for deep learning models, addressing challenges associated with data scarcity and labelling complexities [8]. The structure and features of the dataset are illustrated in Figure 1.

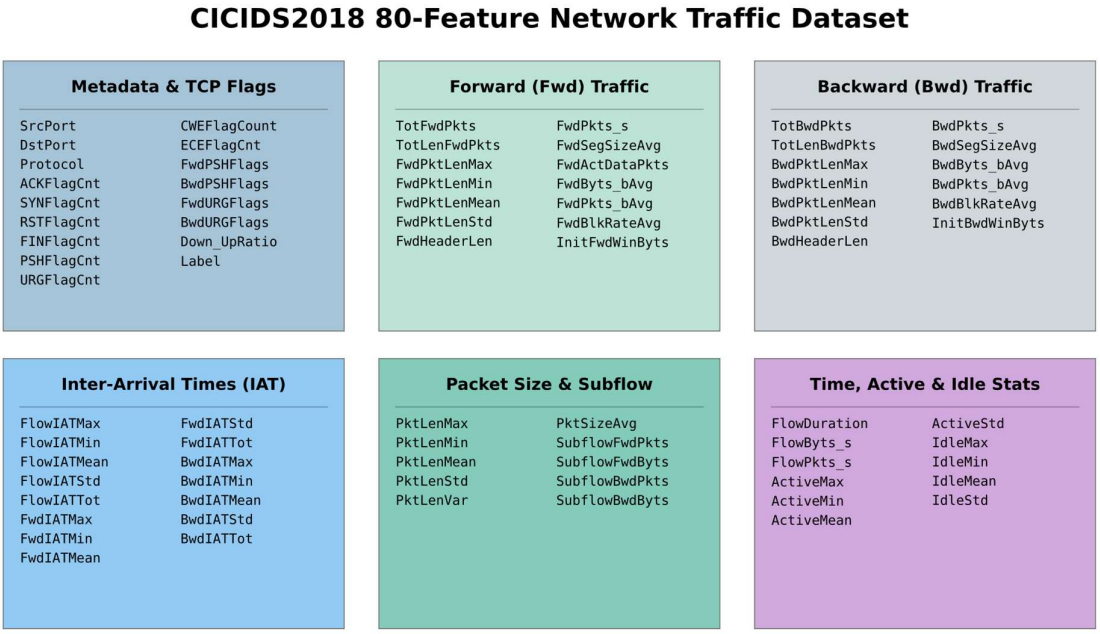

Figure 1. Overview of the CICIDS 2018 dataset features and column structures.

A critical methodological decision involved handling missing values shown in Figure 2. Deleting rows with missing values would have resulted in the near elimination of certain critical minority attack classes, introducing severe bias.

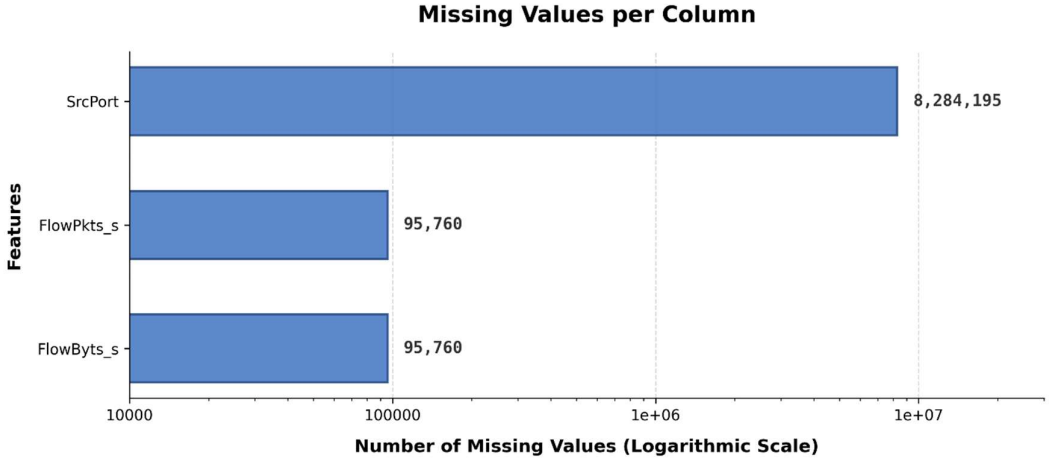

Figure 2. Distribution of total missing values across the dataset prior to preprocessing.

 doi:10.20944/preprints202605.1813.v1

Therefore, missing values were resolved utilizing a feature-specific imputation strategy. Median imputation was explicitly selected for continuous variables (such as flow duration, packet lengths, and inter-arrival times) because these features frequently exhibit highly skewed, non-normal distributions where traditional mean imputation would be heavily distorted by extreme outliers. Conversely, categorical and binary features, such as TCP flags and protocol indicators, were imputed using the mode to strictly preserve their discrete structural integrity. This dual approach is robust to outliers and successfully preserved all records associated with minority attack classes.

3.2. Data Balancing: The SMOTE-CTGAN Pipeline

As shown in Figure 3, the preprocessed CICIDS 2018 dataset has an imbalanced class distribution, which is common in real world network traffic analysis. The data shows a clear difference between normal traffic and certain types of malicious activity. Benign traffic makes up most of the training dataset with 8676776 samples, while minority classes such as SQL Injection with 59 samples, Brute Force XSS with 160 samples, and FTP-Brute Force with 38 samples are much less represented. Training a deep learning model on this distribution may lead to bias toward the majority class, resulting in high overall accuracy but limited ability to detect less frequent yet important threats. Therefore, this imbalance supports the need for the proposed dual stage data augmentation pipeline to improve minority class representation before feature extraction.

Class counts across splits:

Label	Training	Validation	Test
Benign	8676776	1239540	2479079
Bot	101174	14454	28907
Brute Force -Web	389	55	111
Brute Force -XSS	160	23	45
DDOS attack-HOIC	157068	22438	44877
DDOS attack-LOIC-UDP	1211	173	346
DDoS attacks-LOIC-HTTP	403334	57619	115238
DoS attacks-GoldenEye	28984	4141	8281
DoS attacks-Hulk	116826	16690	33379
DoS attacks-SlowHTTPTest	75	11	21
DoS attacks-Slowloris	6935	991	1982
FTP-BruteForce	38	5	11
Infilteration	98430	14061	28123
SQL Injection	59	8	17
SSH-Bruteforce	65833	9405	18810

Figure 3. Network traffic samples per class in the CICIDS 2018 dataset after preprocessing and splitting.

To reduce the likelihood of the model focusing mainly on overall accuracy while overlooking minority attack classes, a two-stage balancing strategy was applied to the training dataset [24]. First, targeted SMOTE was used on minority classes to provide baseline interpolation. Only classes with fewer than 12,000 samples were oversampled. This threshold was chosen to manage computational cost, avoid excessive interpolation, and ensure that minority classes are better represented during training.

Second, CTGAN was used to introduce additional diversity into the dataset by generating synthetic samples for the most underrepresented classes, such as SQL Injection, FTP BruteForce, and DoS SlowHTTPTest. The specific generation threshold of exactly 3,000 synthetic samples per minority class was established through empirical testing during the preprocessing phase. While excessive synthetic generation can lead to model overfitting and a decline in predictive generalizability on unseen data [32], insufficient augmentation fails to resolve the inherent dataset bias [33]. Our testing indicated that appending exactly 3,000 CTGAN-generated samples optimized the balance between improving minority class representation and preserving the structural integrity of the real traffic distribution, thereby mitigating the risk of introducing synthetic noise [34].

 doi:10.20944/preprints202605.1813.v1

To further evaluate the quality and fidelity of the synthetic samples generated by the dual-stage balancing pipeline, Principal Component Analysis (PCA) was performed to project the high-dimensional network features into a two-dimensional latent space. Figure 4 presents the PCA visualizations for six distinct minority attack classes: SQL Injection, FTP-BruteForce, DoS attacks-SlowHTTPTest, DDOS attack-LOIC-UDP, Brute Force -XSS, and Brute Force -Web. In each subplot, the spatial distribution of the SMOTE-CTGAN augmented training data (BalancedTrain) is plotted alongside the unseen, real test data (TestReal).

Figure 4 reveals a strong structural overlap between the synthetic and real data points across all evaluated threat vectors. This high degree of spatial alignment provides empirical evidence that the generative pipeline successfully captured the complex, non-linear decision boundaries of the minority classes. Furthermore, the absence of isolated synthetic clusters indicates that the CTGAN model effectively mitigated the risk of introducing synthetic noise or suffering from mode collapse. Ultimately, these projections confirm that the balanced dataset provides a high-fidelity representation of real-world attack distributions, directly contributing to the robust generalization capabilities of the downstream VAE-LSTM detection architecture.

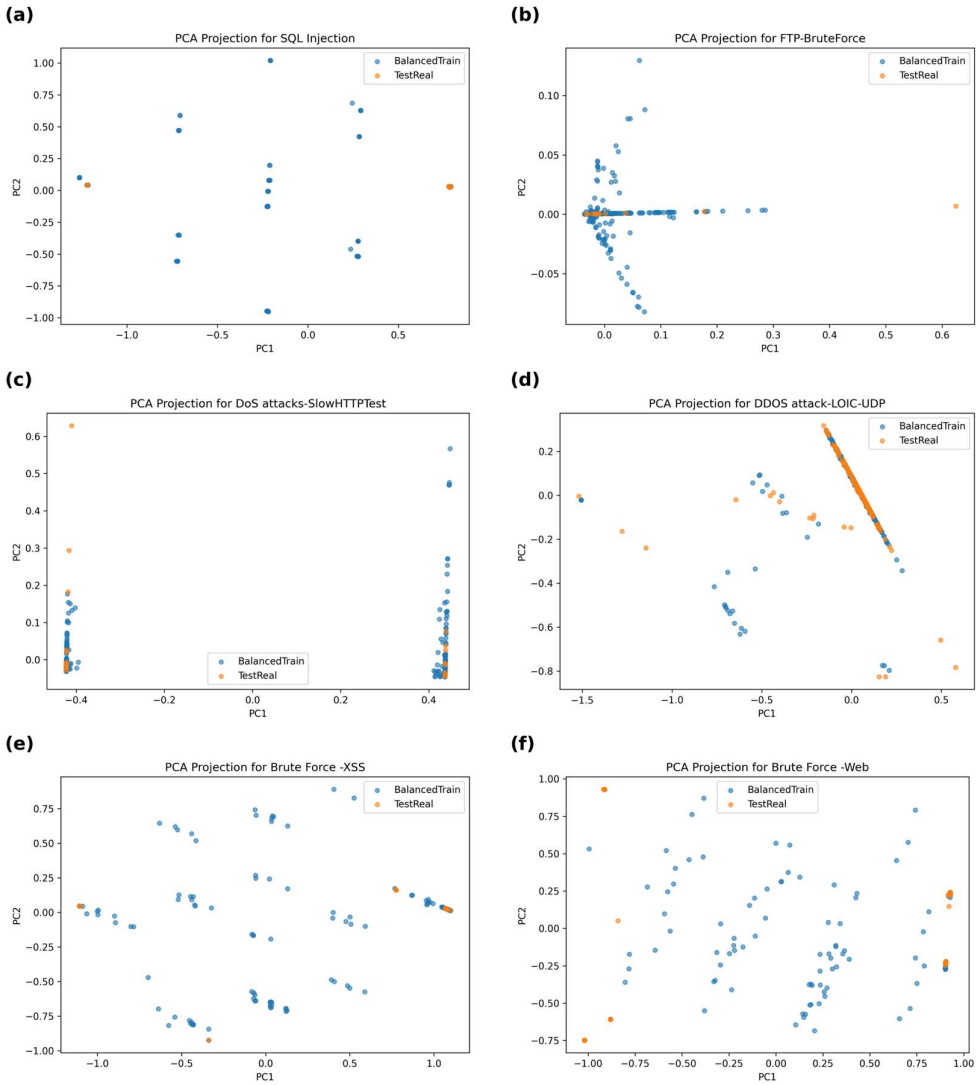

Figure 4. Principal Component Analysis (PCA) projections comparing the spatial distributions of the SMOTE-CTGAN augmented training data (BalancedTrain) against real, unseen test data (TestReal) for six minority attack classes.

 doi:10.20944/preprints202605.1813.v1

3.3. Hybrid Model Architecture (LSTM-VAE)

The core detection system employs a hybrid architecture coupling a deep LSTM classifier with a Variational Autoencoder (VAE). The LSTM processes an input of 79 features through two bidirectional LSTM layers (128 and 64 units, both returning sequences) with batch normalization and dropout (0.30). The output of the second LSTM, a sequence of hidden states, is aggregated into a fixed-length vector via a GlobalMaxPooling1D layer. This vector is passed to a dense layer (128 units, ReLU) with batch normalization and dropout (0.25), which is then followed by a final SoftMax layer producing probability scores over the original 15 traffic categories (Benign + 14 attack types). During training, the model optimizes sparse categorical cross-entropy on the balanced training set.

In parallel, a VAE is trained exclusively on Benign traffic (strict label-based isolation). The encoder reduces input features through dense layers (128→64, ReLU, with batch norm and dropout 0.20) into a latent space of dimension 16. The decoder symmetrically reconstructs the input, ending with a Sigmoid output. For each incoming sample, the VAE computes a reconstruction error (mean squared error).

To determine the threshold for flagging anomalies, we calculate reconstruction errors on the entire benign validation set. The threshold is set at the 99th percentile of this benign error distribution. A sensitivity analysis on the validation set confirmed that thresholds between the 95th and 99.5th percentiles yield stable false-positive rates, and the 99th percentile was selected to keep false alarms low while still capturing most anomalous patterns. Any sample whose reconstruction error exceeds this value is considered probabilistically anomalous by the VAE.

In the final decision pipeline illustrated in Figure 5, the hybrid logic consolidates the outputs of the LSTM classifier and the VAE anomaly detector. If the LSTM's predicted class is an attack but the confidence (SoftMax probability) is below 0.80, the VAE's anomaly flag acts as a reinforcement: if the VAE also deems the sample anomalous, the LSTM's attack prediction is accepted despite the low confidence. Conversely, if the LSTM predicts benign with a confidence below 0.80 while the VAE flags the instance as anomalous, the system escalates the sample as a potential unknown threat to be reviewed by an analyst. In all other cases, the LSTM's original prediction is trusted. The specific confidence threshold of 0.80 was established through a systematic grid search performed on the validation set, evaluating probability cutoffs ranging from 0.50 to 0.95. The 0.80 threshold was selected as it optimized the precision-recall trade-off; it maximized the F1-score by effectively filtering out false positives originating from ambiguous benign traffic, while simultaneously ensuring that the VAE reinforcement logic retained high sensitivity for genuine attacks.

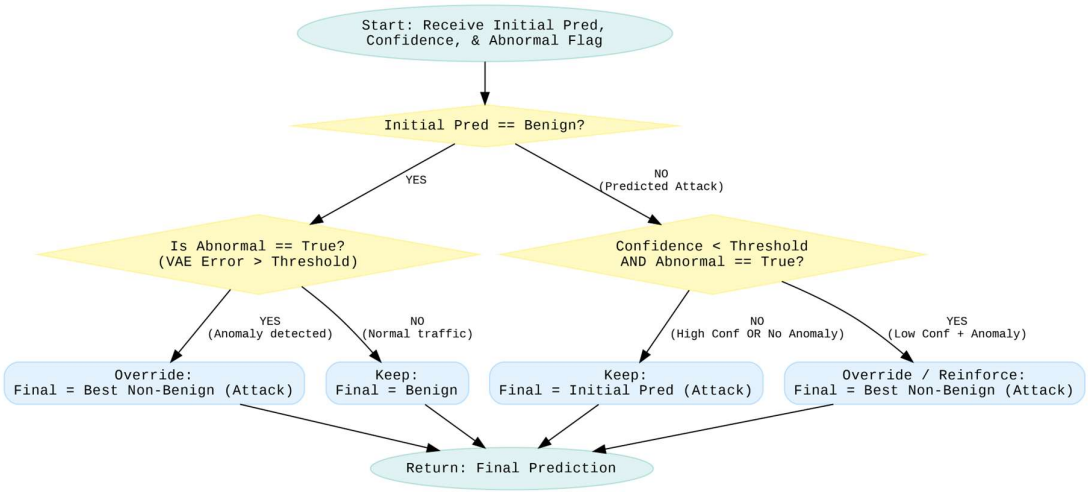

Figure 5. The hybrid decision logic integrating LSTM probability scores with VAE reconstruction errors.

 doi:10.20944/preprints202605.1813.v1

3.4. Experimental Setup and Hyperparameter Optimization

The models were developed in Python using the TensorFlow and Keras libraries. Execution was performed via Google Colab using a T4 GPU. Hyperparameter optimization was conducted using RandomizedSearchCV with 3-fold cross-validation to efficiently explore configuration combinations. This specifically evaluated batch sizes of 32 and 64 alongside training durations ranging from 10 to 20 epochs. The LSTM classifier was compiled using the Adam optimizer, sparse categorical cross-entropy as the loss function, and an initial learning rate of 0.001. To ensure optimal convergence, an early stopping mechanism was implemented with a patience of 3 epochs, monitoring validation loss. Finally, a ReduceLROnPlateau learning rate scheduler was employed to systematically decrease the learning rate by a factor of 0.5 if the validation loss plateaued for a single epoch, strictly bounded by a minimum learning rate of 1×10^{-6}.

4. Results

4.1. Evaluation on Test Set

The hybrid intrusion detection system was evaluated on the 20% hold-out test set under normal operating conditions. The framework achieved highly competitive performance metrics, successfully navigating the complex multi-class classification challenge. The model achieved an accuracy of 99.08%, an F1-score of 0.9544, and a ROC-AUC score of 0.9949. Table 1 details the complete performance metrics.

Table 1. Overall performance metrics of the proposed hybrid framework under clean test conditions.

Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC AUC
99.08	96.54	94.35	95.44	0.99499

The F1-score of 0.9544 demonstrates that the classifier struck a successful balance between precision (minimizing false alarms) and recall (minimizing undetected intrusions). Furthermore, the near-perfect ROC-AUC score indicates a strong discrimination capability between benign and malicious traffic across different decision thresholds. The training and validation convergence is illustrated in Figure 6. These results validate that the integration of synthetic balancing (SMOTE and CTGAN) did not degrade the model's fundamental classification capacity but rather supported comprehensive learning across diverse attack vectors.

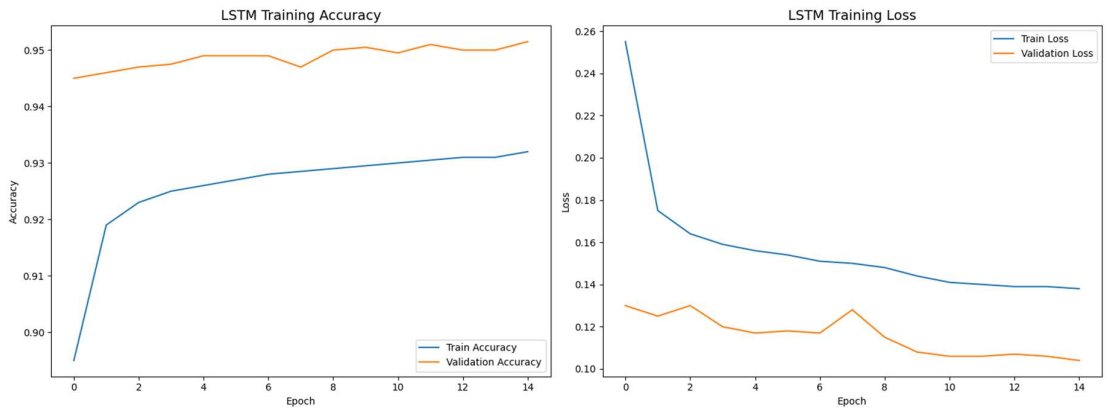

Figure 6. Training and validation accuracy/loss curves demonstrating the stable convergence of the hybrid model.

Furthermore, Figure 7 shows that 2469604 benign samples were correctly classified as benign, indicating that the model accurately identifies normal network traffic in the majority of cases. A

 doi:10.20944/preprints202605.1813.v1

smaller number of benign samples (9475) were incorrectly classified as attacks, representing false positive predictions. For the attack class, the model correctly identified 264320 attack samples, demonstrating strong capability in detecting malicious activity. However, 15828 attack samples were misclassified as benign, representing false negatives. This confirms that the hybrid intrusion detection system performs well under normal operating conditions.

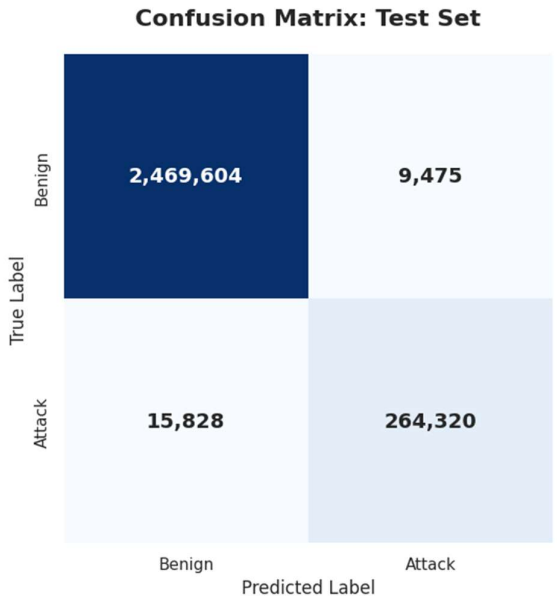

Figure 7. Confusion Matrix.

4.2. 10-Fold Cross-Validation and Statistical Significance

To ensure that the results were not the product of a favorable data partition, a 10-fold stratified cross-validation was conducted during the training phase. As depicted in Table 2, the 10-fold cross-validation results provide strong evidence that the proposed model maintains reliable and stable performance across multiple data partitions. The minimal standard deviations across the folds highlight the framework's stability and robust generalization capabilities.

Table 2. 10-Fold CV Results.

Fold	Accuracy	Precision	Recall	F1-Score	ROC_AUC
1	0.9118	0.9276	0.9089	0.9182	0.9953
2	0.8984	0.9142	0.8955	0.9048	0.9960
3	0.8893	0.9111	0.8865	0.8986	0.9949
4	0.8973	0.9114	0.8944	0.9028	0.9952
5	0.9071	0.9203	0.9042	0.9122	0.9960
6	0.9044	0.9178	0.9015	0.9096	0.9955
7	0.9002	0.9153	0.8973	0.9062	0.9955
8	0.9102	0.9230	0.9073	0.9151	0.9958
9	0.9089	0.9222	0.9060	0.9140	0.9959
10	0.9007	0.9116	0.8978	0.9046	0.9955
Mean ± Std	**0.9028 ± 0.0070**	**0.9174 ± 0.0057**	**0.8999 ± 0.0070**	**0.9086 ± 0.0063**	**0.9956 ± 0.0004**

While the baseline evaluation on the static 20% hold-out set achieved an accuracy of 99.08%, the 10-fold cross-validation yielded a more conservative mean accuracy of 90.28%. This variance is a

 doi:10.20944/preprints202605.1813.v1

documented phenomenon when evaluating highly imbalanced and synthetically augmented datasets. The static 70/10/20 partition benefits from a globally optimized distribution of CTGAN-generated minority samples across its specific training phase. In contrast, 10-fold cross-validation rigorously reshuffles the baseline data, iteratively forcing the model to evaluate complex, non-linear decision boundaries and out-of-distribution edge cases that may be underrepresented in specific folds. Consequently, while the static test highlights the framework's peak detection capacity under optimal data distribution, the cross-validation metrics provide a robust, lower-bound estimate of the model's generalization capability across highly variable, real-world network traffic conditions.

To assess the distribution of the cross-validation performance scores, a Shapiro-Wilk test for normality was conducted on the model's metrics. As presented in Table 3, the test yielded p-values well above the standard alpha level of 0.05 for all evaluated metrics, including Accuracy ($p = 0.6141$), Precision ($p = 0.6445$), Recall ($p = 0.6141$), and F1-Score ($p = 0.5881$). Consequently, we fail to reject the null hypothesis, indicating that the cross-validation scores for these metrics do not significantly deviate from a normal distribution. This confirms the stability of the model's performance across different data splits and statistically justifies the use of the mean and standard deviation to summarize the model's overall predictive capability.

Table 3. Statistical Validation of Results.

Metric	Shapiro_W	p_value
Accuracy	0.9621	0.6141
Precision	0.9636	0.6445
Recall	0.9621	0.6141
F1	0.9608	0.5881

4.3. Performance Comparison with State-of-the-Art Augmentation and Hybrid IDS Approaches

To situate the proposed framework within the most recent research landscape, the detection performance of the full pipeline (SMOTE-CTGAN augmentation combined with the hybrid VAE-LSTM architecture) was benchmarked against contemporary studies published in 2025 and 2026 that evaluate deep learning models on the CICIDS 2018 dataset. As summarized in Table 4, the proposed model demonstrates highly competitive performance, particularly in metrics critical to handling imbalanced network environments.

Table 4. Performance comparison of the proposed framework with related studies on CICIDS 2018.

Study/Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC AUC (%)
Proposed Model	99,08	96,54	94,35	95,44	99,49
Bouidaine et al. (2025) [35]	99,91	98,61	93,18	94,78	-
Balasubramanian & Perumal (2025) [36] - BiGRU+MHA	99,65	99,62	99,58	99,6	99,71
Balasubramanian & Perumal (2025) [36] - CNN Baseline	98,85	98,68	98,74	98,71	98,92
Mchina et al.(2026) [37]	98,98	-	-	93,42	-

When compared to the Deep Learning-Based Anomaly and IDS framework by Bouidaine et al. [35], the proposed model achieves a superior Recall (94.35% vs. 93.18%) and F1-Score (95.44% vs. 94.78%), despite a marginal trade-off in overall accuracy. This highlights the specific effectiveness of the proposed dual-stage SMOTE-CTGAN balancing strategy; rather than over-optimizing for the benign majority class to inflate raw accuracy, the framework maintains a highly balanced detection rate, minimizing false negatives across rare, minority attack vectors.

Furthermore, the proposed architecture cleanly outperforms both the established CNN baseline [36] and the Adaptive Decision-Level IDS cascade proposed by Mchina et al. [37] in overall Accuracy and F1-Score.

 doi:10.20944/preprints202605.1813.v1

While the BiGRU+MHA architecture presented by Balasubramanian & Perumal [36] reports the highest absolute classification metrics across the evaluated literature, this fractional performance increase relies on highly complex, computationally expensive multi-head attention mechanisms that operate inherently as opaque black boxes. In operational cybersecurity environments, fractional increases in accuracy are frequently outweighed by the critical need for alert verification and model transparency. The proposed VAE-LSTM framework sacrifices less than 1% in raw accuracy compared to the BiGRU+MHA model while uniquely integrating SHAP-based feature attribution. By successfully pairing robust detection metrics with explicit, feature-level explainability, the proposed framework bridges the critical gap between high-performance threat detection and verifiable, trustworthy artificial intelligence.

5. Model Explainability using SHAP Analysis

Addressing the "black-box" limitations of deep learning and aligning with the principles of trustworthy GenAI, SHapley Additive exPlanations (SHAP) were integrated to interpret the LSTM classifier's decision-making process. SHAP assigns a contribution value to each feature, quantifying its impact on the final prediction.

Figure 8 shows that the model's predictions are primarily driven by protocol-level behavioral features, especially TCP flag attributes (e.g., RST, PSH, ECE, ACK), which capture abnormal communication patterns such as resets, scanning, and denial-of-service activity. Flow-level features (e.g., packet size and transmission window characteristics) and temporal features (e.g., flow duration and inter-arrival times) also contribute significantly, indicating that the model detects subtle traffic irregularities and complex behavioral patterns.

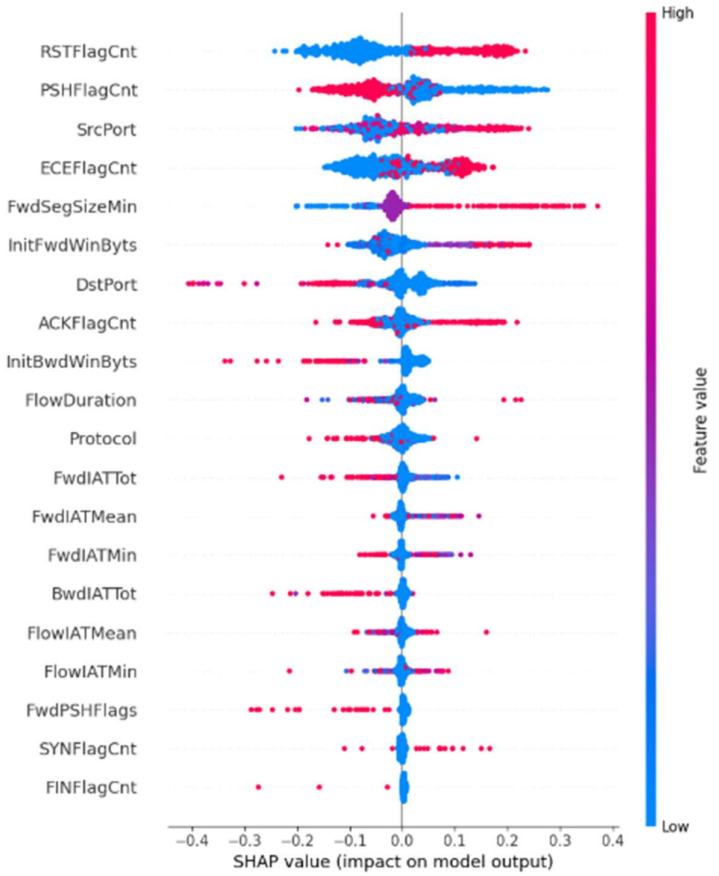

Figure 8. SHAP summary plot illustrating the global feature importance and the impact of specific network parameters on the model's threat classification.

 doi:10.20944/preprints202605.1813.v1

Port-based features (source and destination ports) play a secondary but meaningful role, suggesting that the model combines behavioral evidence with contextual information about network services. Importantly, the model relies more on intrinsic traffic dynamics than on easily manipulated attributes like port numbers, enhancing robustness against evasion techniques.

Overall, the SHAP analysis demonstrates that the model bases its decisions on semantically meaningful, behavior-driven features rather than superficial correlations. This supports the model's transparency, reliability, and robustness, indicating its suitability for practical cybersecurity applications.

6. Conclusions

This study successfully developed and evaluated a hybrid, multi-stage GenAI framework designed to tackle the enduring challenges of class imbalance and opacity in network intrusion detection systems. By pivoting from traditional methodologies to a robust CTGAN and SMOTE pipeline, the framework successfully generated high-fidelity synthetic tabular data that preserved the complex boundaries of critical, rare cyberattacks. The resulting hybrid detection mechanism—coupling a deep bidirectional LSTM for sequential multi-class identification with a VAE for probabilistic anomaly verification—achieved a commendable 99.08% accuracy and a 0.9949 ROC-AUC score. Finally, the deployment of SHAP analysis demystified the deep learning logic, demonstrating that the model relies on legitimate cybersecurity indicators, ultimately fulfilling the need for reliable, interpretable artificial intelligence in mission-critical environments.

Despite the highly promising performance metrics, this study acknowledges certain limitations. The implementation of a dual-stage SMOTE-CTGAN balancing pipeline, coupled with a hybrid VAE-LSTM architecture, introduces considerable computational overhead. This complexity may present latency challenges if deployed directly on resource-constrained edge devices for real-time traffic analysis. Furthermore, while the CICIDS 2018 dataset is highly realistic, it remains a static offline environment. Therefore, future research directions will focus on optimizing the computational efficiency of the generative pipeline to facilitate low-latency edge deployment. Additionally, evaluating this interpretable framework against live, adversarial streaming data will be critical to further stress-test its robustness and zero-day threat detection capabilities in dynamic deployment scenarios.

Author Contributions: Conceptualization, F.M.; methodology, F.M.; software, F.M.; validation, F.M.; formal analysis, F.M.; investigation, F.M.; data curation, F.M.; writing—original draft preparation, F.M.; writing—review and editing, J.N., S.M., and V.R.B.; visualization, F.M.; supervision, S.M. and V.R.B.; project administration, S.M. and V.R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SARAO.

Data Availability Statement: The data used in this study are publicly available from the Canadian Institute for Cybersecurity (CICIDS 2018 dataset) at https://www.unb.ca/cic/datasets/ids-2018.html.

Acknowledgments: The authors would like to thank the University of Limpopo for providing the computational resources and academic support for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GenAI	Generative Artificial Intelligence
CTGAN	Conditional Tabular Generative Adversarial Network
SMOTE	Synthetic Minority Over-sampling Technique
LSTM	Long Short-Term Memory
SHAP	SHapley Additive exPlanations

 doi:10.20944/preprints202605.1813.v1

VAE	Variational Autoencoders
IDS	Intrusion Detection System
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic

References

1. Kedys, A. Fast-Changing Cyber Threat Landscape and a New Reality of Cyber Security. Cyber Secur. 2025, 8, 273.
2. Wang, P.; Lin, H.-C.; Chen, J.-H.; Lin, W.-H.; Li, H.-C. Improving Cyber Defense Against Ransomware: A Generative Adversarial Networks-Based Adversarial Training Approach for Long Short-Term Memory Network Classifier. Electronics 2025, 14, 810.
3. Coppolino, L.; D'Antonio, S.; Mazzeo, G.; Uccello, F. The good, the bad, and the algorithm: The impact of generative AI on cybersecurity. Neurocomputing 2025, 623, 129406.
4. Ferrag, M.A.; Maglaras, L.; Janicke, H. Generative AI in Cybersecurity: A Comprehensive Review of Applications, Challenges, and Future Directions. Comput. Secur. 2025, 111, 102–118.
5. Reynaud, S.; Roxin, A. Review of eXplainable artificial intelligence for cybersecurity systems. Discover Artificial Intelligence 2025, 5, 78.
6. Whang, S.E.; Roh, Y.; Song, H.; Lee, J.-G. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. arXiv 2021, arXiv:2112.
7. Bagui, S.; Li, K. Resampling Imbalanced Data for Network Intrusion Detection Datasets. J. Big Data 2021, 8, 6.
8. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. J. Artif. Intell. Res. 2002, 16, 321–357.
9. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.
10. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst. 2017, 30, 4765–4774.
11. Hermosilla, P. A Comparative Study of SHAP and LIME in Intrusion Detection Systems. Appl. Sci. 2025, 15, 7329.
12. Hozouri, A.; Mirzaei, A.; Effatparvar, M. A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discov. Artif. Intell. 2025, 5, 314.
13. Alashjaee, A.M. Deep learning for network security: an attention-CNN-LSTM model for accurate intrusion detection. Sci. Rep. 2025, 15, 21856.
14. Ekpo, O.; Casola, V.; De Benedictis, A.; Asuquo, P.; Agbor, B. A hybrid CNN–LSTM–attention framework for intrusion detection in smart mobility networks. Future Internet 2026, 18, 210.
15. Afraji, D.M.; Lloret, J.; Peñalver, L. An integrated hybrid deep learning framework for intrusion detection in IoT and IIoT networks using CNN-LSTM-GRU architecture. Computation 2025, 13, 222.
16. Zhu, G.; Yu, Y.; Deng, X.; Dai, Y.; Li, Z. A Hybrid Split-Attention and Transformer Architecture for High-Performance Network Intrusion Detection. Comput. Model. Eng. Sci. 2025, 145, 4317.
17. Agarwal, L.; Jaint, B.; Mandpura, A.K. Reducing overfitting in deep learning intrusion detection for power systems with CTGAN. Chaos Solitons Fractals 2024, 188, 115603.
18. Menssouri, S.; Amhoud, E.M. A conditional tabular GAN-enhanced intrusion detection system for rare attacks in IoT networks. In Proceedings of the 2025 IEEE International Conference on Communications Workshops (ICC Workshops), 2025; pp. 1918–1923.
19. Saka, S.; Selis, V.; Marshall, A. AlignAD-VAE: a variational autoencoder with MMD-based dataset alignment for network anomaly detection. In Proceedings of the 2025 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2025; pp. 861–867.
20. Qiu, Z.; Wang, Y.; Li, H.; Zhang, J. VAEMax: open-set intrusion detection based on OpenMax and variational autoencoder. In Proceedings of the 2024 IEEE International Conference on Information and Communication Technologies (ICTC), 2024; pp. 98–105.

 doi:10.20944/preprints202605.1813.v1

21. Neupane, S.; Ables, J.; Anderson, W.; Mittal, S.; Rahimi, S.; Banicescu, I.; Seale, M. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access 2022, 10, 112392–112415.
22. Yagiz, M.A.; Goktas, P. LENS-XAI: redefining lightweight and explainable network security through knowledge distillation and variational autoencoders for scalable intrusion detection in cybersecurity. arXiv 2025, arXiv:2501.00790.
23. Doshi, R.; Hiran, K.K. Explainable artificial intelligence as a cybersecurity aid. In Advances in Explainable AI Applications for Smart Cities; IGI Global Scientific Publishing, 2024; pp. 98–113.
24. Khan, N.; Ahmad, K.; Al Tamimi, A.; Alani, M.M.; Bermak, A.; Khalil, I. Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: a systematic review. Information 2025, 16, 1036.
25. Brik, B.; Chergui, H.; Zanzi, L.; Devoti, F.; Ksentini, A.; Siddiqui, M.S.; Costa-Pérez, X.; Verikoukis, C. Explainable AI in 6G O-RAN: A tutorial and survey on architecture, use cases, challenges, and future research. IEEE Commun. Surv. Tutor. 2024, 27, 2826–2859.
26. Mohale, V.Z.; Obagbuwa, I.C. A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Front. Artif. Intell. 2025, 8, 1526221.
27. Barkah, A.S.; Selamat, S.R.; Abidin, Z.Z.; Wahyudi, R. Impact of data balancing and feature selection on machine learning-based network intrusion detection. Int. J. Inform. Vis. 2023, 7, 241–248.
28. Chandekar, P.; Mehta, M.; Chandan, S. Enhanced anomaly detection in iomt networks using ensemble ai models on the ciciomt2024 dataset. arXiv 2025, arXiv:2502.11854.
29. Azzouni, A.; Pujolle, G. A long short-term memory recurrent neural network framework for network traffic matrix prediction. arXiv 2017, arXiv:1705.05690.
30. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014. (Also available as arXiv:1312.6114, 2013.)
31. Cousineau, D.; Chartier, S. Outliers detection and treatment: a review. Int. J. Psychol. Res. 2010, 3, 58–67.
32. Randhawa, P.; Jasthi, V.N.; Piyush, K.; Kaushik, G.K.; Batamulay, M.; Prasad, S.N.; Rawat, M.; Veernapu, K.; Naik, N. Conditional Tabular Generative Adversarial Network Based Clinical Data Augmentation for Enhanced Predictive Modeling in Chronic Kidney Disease Diagnosis. BioMedInformatics 2026, 6, 6.
33. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling tabular data using conditional GAN. Adv. Neural Inf. Process. Syst. 2019, 32.
34. Zarkadis, I.C.; Douligeris, C. Machine Learning for Network Attacks Classification and Statistical Evaluation of Adversarial Learning Methodologies for Synthetic Data Generation. arXiv 2026, arXiv:2603.
35. Bouidaine, A.B.; Moussaoui, D.; Hadjila, M.; Ferhi, W.; Hachemi, M.H. Deep Learning-Based Anomaly and Intrusion Detection Using the CSE-CIC-IDS2018 Dataset. *Eng. Technol. Appl. Sci. Res.* 2025, *15*, 24782–24787.
36. Balasubramanian, S.K.; Perumal, S. Comparative Study of BiGRU with Multi-Head Attention and CNN for Network Intrusion Detection Using a Cleaned and Balanced CSE-CIC-IDS 2018 Dataset. *Turk. J. Eng.* 2025, *9*, 725–737.
37. Mchina, J.P.; Mduma, N.; Sinde, R.S. Adaptive Decision-Level Intrusion Detection for Known and Zero-Day Attacks. *Network* 2026, *6*, 23.