Explainable Zero-Day Intrusion Detection Using SHAP-Enhanced Deep Learning on the CICIDS2017 Dataset

Obeten O. Micheal[1], Ahmed Jimoh[2]
Department of Science Laboratory Technology
Auchi Polytechnic Auchi Edo State Nigeria
Corresponding author's email: michealobeten@auchipoly.edu.ng[1]

Abstract— Nigerian research institutions operate under severe energy and financial constraints, with laboratories consuming disproportionate shares of institutional budgets while generating significant environmental loads. Green laboratory technology (GLT) adoption offers a dual pathway to address both operational costs and ecological sustainability, yet empirical quantification of these benefits within the Nigerian context remains sparse. This paper presents a comprehensive analysis of the dual impact of GLT adoption on cost reduction and environmental sustainability in Nigerian research institutions, synthesising global evidence and contextualising findings for the Nigerian operating environment. A systematic narrative review was conducted drawing on peer-reviewed literature from Scopus, PubMed, Web of Science, and Google Scholar. Sources were screened for relevance to laboratory sustainability, energy economics, waste management, and institutional practice in Africa and comparable developing-country contexts. Thirty high-quality references were synthesised. Evidence indicates that certified green laboratory programmes reduce energy consumption by 20–50%, lower operational costs by up to $39,000 per academic laboratory annually, and cut carbon dioxide-equivalent (CO_2e) emissions by an average of 31.32 tonnes per laboratory per year. In the Nigerian context, solar photovoltaic microgrid integration has been shown to reduce energy costs substantially and improve research reliability across universities. Chemical waste mismanagement in African academic laboratories poses significant public health and environmental risks that GLT addresses directly. GLT adoption is technically feasible and economically justifiable for Nigerian research institutions. A phased implementation framework anchored by policy reform, institutional capacity building, and green financing instruments is recommended. The dual fiscal and ecological returns position GLT as a strategic imperative rather than an optional enhancement.

Keywords— *green laboratory technology; environmental sustainability; operational cost reduction; Nigerian research institutions; renewable energy; waste management; laboratory efficiency*

I. INTRODUCTION

Network security threats are escalating in both frequency and sophistication. According to the Check Point Research Cyber Security Report 2023, global cyber attacks increased by 38% in 2022 compared to the prior year, with enterprise networks facing an average of 1,168 weekly attack attempts [1]. Among the most consequential category of threats are zero-day attacks, defined as exploits that target previously unknown or unpatched vulnerabilities for which no defensive signature yet exists. The Google Threat Intelligence Group tracked 75 actively exploited zero-day flaws in 2024, each representing a window of opportunity during which conventional signature-based defences were entirely blind [2]. Intrusion Detection Systems form a critical layer of network defence, monitoring traffic flows and system events for evidence of malicious activity. Signature-based IDS, the dominant deployed paradigm, match observed traffic against a library of known attack patterns. This approach is highly effective against known threats but is fundamentally incapable of detecting novel attacks whose signatures have not yet been characterised and disseminated [3]. Anomaly-based IDS offer a complementary and more generalisable approach: by learning statistical representations of normal network behaviour, they can in principle detect deviations that correspond to previously unseen attacks, including zero-day exploits [4].

Deep learning has emerged as the leading technical paradigm for anomaly-based IDS, offering the capacity to learn complex, high-dimensional representations of network traffic that capture non-linear patterns beyond the reach of classical machine learning algorithms [5]. Convolutional Neural Networks (CNNs) extract local spatial features from traffic feature vectors; Long Short-Term Memory networks (LSTMs) model temporal dependencies in sequential flow data; and hybrid Conv1D-BiLSTM architectures combine both capacities. Studies on the CICIDS2017 benchmark dataset have reported deep learning detection accuracies exceeding 99% under controlled experimental conditions [6, 7]. Despite their technical performance, deep learning models present a critical barrier to operational adoption: their black-box character prevents security analysts from understanding why a particular prediction was made, undermining trust, making false positive investigation inefficient, and conflicting with regulatory accountability requirements including GDPR and ISO 27001 [8]. Explainable Artificial Intelligence (XAI) frameworks, specifically SHapley Additive exPlanations (SHAP), address this barrier by providing mathematically grounded, feature-level explanations of individual model predictions derived from cooperative game theory [9].

This paper makes three contributions. First, it proposes a Conv1D-BiLSTM deep learning architecture optimised for multi-class intrusion detection on CICIDS2017. Second, it integrates SHAP DeepExplainer to generate both global feature importance rankings and local instance-level explanations that support security operations centre workflows. Third, it evaluates zero-day generalisation capability through a systematic leave-one-attack-out experimental protocol, providing empirical evidence on the framework's ability to detect attack classes unseen during training.

II. RELATED WORK

A. *Deep Learning for Intrusion Detection on CICIDS2017*

The CICIDS2017 dataset, generated by the Canadian Institute for Cybersecurity at the University of New Brunswick, has become the most widely used benchmark for IDS research. Comprising 2,830,743 network flow records described by 80 raw features across 11 attack types and a benign class, it captures realistic traffic patterns including DoS, DDoS, port scanning, brute force, and infiltration attacks [6]. A comprehensive benchmarking study evaluated 10 supervised and unsupervised machine learning algorithms on CICIDS2017, reporting accuracy metrics across 31 model configurations and establishing performance baselines against which subsequent deep learning studies are compared [7].

A comparative study of deep learning and machine learning methods for intrusion detection on CICIDS2017 found that hybrid models combining CNN feature extraction with LSTM temporal modelling consistently outperformed standalone algorithms, with the CNN-LSTM hybrid achieving 97.3% detection accuracy against DDoS attacks [10]. A deep learning and machine learning ensemble approach applied to CICIDS2017 achieved an average inference time of 2.4 milliseconds per instance on standard hardware, demonstrating real-time viability for security operations centre deployment [11]. A two-stage XAI-enhanced framework designated XI2S-IDS achieved 99.81% accuracy on CICIDS2017 while using SHAP to provide global explanations for the binary detection classifier, validating the compatibility of high accuracy and model transparency [12].

B. *Explainable AI for Intrusion Detection*

The integration of XAI techniques into IDS represents one of the most active research frontiers in network security. A study designing an explainable IDS with multi-model architecture incorporating Random Forest, SVM, and DNN alongside SHAP and LIME explanations demonstrated that strategic XAI integration transforms IDS from opaque alert generators into collaborative defence tools, enabling human-AI teamwork against evolving cyber threats [8]. A study detecting cybersecurity threats by integrating SHAP interpretability and strategic data sampling demonstrated that SHAP-based global and local explanations provide both model-level insight and per-prediction justification that align with security analyst operational workflows [13].

A versatile XAI-based IDS framework using ANOVA-based feature selection combined with global SHAP scores retained only the ten most informative features, achieving approximately 70% dimensionality reduction while maintaining competitive detection accuracy, demonstrating the dual value of SHAP both for explanation and for computationally efficient feature engineering [14]. A systematic review of explainable AI for IDS in Industry 5.0 contexts concluded that SHAP is the most widely adopted XAI technique for network security applications, valued for its theoretical grounding in Shapley values, model-agnostic applicability, and ability to generate both summary plots and local waterfall explanations [15].

C. *Zero-Day Attack Detection*

Zero-day attack detection is a fundamentally more challenging problem than known-attack classification because no labelled examples of the target attack type are available during training. Deep autoencoder architectures that learn compressed representations of normal traffic behaviour and flag anomalies based on reconstruction error have emerged as the dominant technical approach to unsupervised zero-day detection [16]. An intelligent zero-day attack detection system using unsupervised machine learning demonstrated that autoencoder-based anomaly scoring can identify novel attack patterns with meaningful accuracy without requiring labelled attack data [17].

A hybrid zero-day detection framework combining CNN and LSTM networks for IoT security demonstrated that supervised deep learning models retain useful generalisation to novel attack patterns when the feature space of the withheld attack class overlaps with that of known attack types [18]. A review of machine learning-based zero-day attack detection identified anomaly-based IDS, deep autoencoders, and open-set recognition frameworks as the three principal technical lineages, noting that models trained on broad attack taxonomies exhibit greater generalisation to novel patterns than single-attack-type specialists [19]. The integration of SHAP with zero-day detection frameworks, however, remains underexplored: the present study contributes directly to this gap.

III. DATASET AND PREPROCESSING

A. *CICIDS2017 Dataset Description*

The CICIDS2017 dataset was used as the experimental benchmark for this study. Generated by the Canadian Institute for Cybersecurity over five days in July 2017, it contains 2,830,743 network flow records extracted from packet captures using the CICFlowMeter tool [6]. Each record is described by 80 features encompassing forward and backward packet statistics, flow duration, inter-arrival times, flag counts, and payload characteristics. The dataset includes 12 traffic classes: benign traffic and 11 attack categories including DoS Hulk, PortScan, DDoS, DoS GoldenEye, FTP Patator, SSH Patator, DoS Slowloris, DoS Slowhttp, Bot, Web Attacks, and Infiltration.

Figure 1 presents the class distribution of the CICIDS2017 dataset. A severe class imbalance is evident: benign traffic accounts for 2,273,097 records (80.3% of the dataset), while attack classes range from 231,073 records for DoS Hulk to only 36 records for Infiltration. This imbalance is representative of real network environments but creates significant challenges for classifier training, as models trained on imbalanced data tend toward majority-class prediction bias that artificially inflates accuracy while suppressing minority-class recall.

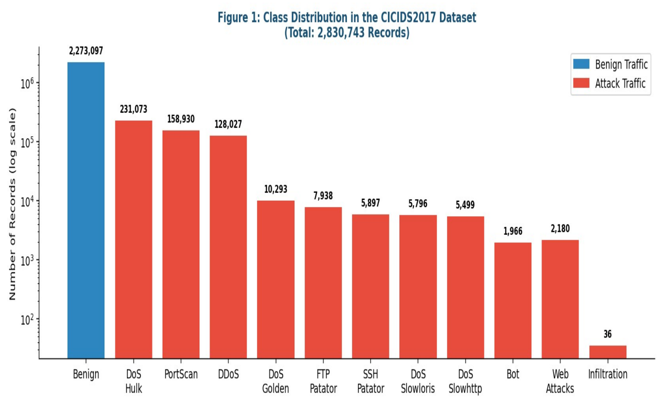

Figure 1: Class Distribution in the CICIDS2017 Dataset Showing Severe Imbalance Between Benign and Attack Classes (Total: 2,830,743 Records)

B. *Data Preprocessing Pipeline*

Preprocessing followed a four-stage pipeline. In the first stage, data cleaning removed 80 duplicate records and 11 features containing infinite or constant values that provide no discriminative information. Two features (Flow Bytes/s and Flow Packets/s) contained infinite values resulting from zero-duration flows and were imputed with the column maximum before scaling. Seventy-eight features were retained after cleaning.

In the second stage, feature scaling was applied using MinMax normalisation to map all feature values to the range 0 to 1, preventing features with large numerical ranges from dominating gradient calculations during neural network training. In the third stage, SMOTE oversampling was applied to the training set only to address class imbalance, generating synthetic minority-class examples through interpolation in feature space rather than simple duplication. SMOTE was applied to all attack classes with fewer than 10,000 records, balancing the training distribution without introducing data leakage into the test set. In the fourth stage, string class labels were encoded as integers for multi-class classification training.

IV. PROPOSED CONV1D-BILSTM MODEL WITH SHAP INTEGRATION

A. *Network Architecture*

The proposed architecture, illustrated in Figure 2, combines one-dimensional convolutional layers for local feature extraction with a Bidirectional LSTM layer for contextual representation learning. The model accepts a 78-dimensional input feature vector, reshaped to a sequence tensor suitable for Conv1D processing. Two Conv1D layers with 64 and 128 filters respectively and kernel size 3 are applied with ReLU activation and batch normalisation between layers. A MaxPooling1D layer with pool size 2 follows for dimensionality reduction.

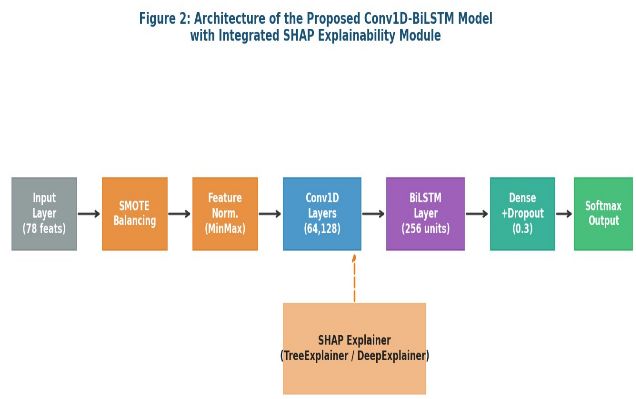

Figure 2: Architecture of the Proposed Conv1D-BiLSTM Model with Integrated SHAP Explainability Module

The pooled output is fed into a Bidirectional LSTM layer with 256 units, which processes the feature representation in both forward and backward temporal directions, capturing bidirectional contextual dependencies in the feature sequence that unidirectional LSTM layers miss. A Dense layer with 128 units and ReLU activation follows, regularised by a Dropout layer with rate 0.3 to prevent overfitting. The output layer applies the Softmax activation function over 12 neurons corresponding to the 12 traffic classes. The model is compiled with the Adam optimiser at a learning rate of 0.001, categorical cross-entropy loss, and early stopping with patience of 10 epochs monitoring validation loss.

B. *SHAP Integration*

SHAP DeepExplainer was integrated with the trained Conv1D-BiLSTM model to generate feature attribution values for both global and local explanation tasks. DeepExplainer, designed specifically for deep neural networks, computes SHAP values by propagating expected value contributions backwards through the network using a background dataset of 1,000 randomly sampled training instances as the reference distribution [9]. Global feature importance was computed as the mean absolute SHAP value across all test set instances, providing a dataset-level ranking of feature contributions. Local instance-level explanations were generated as signed SHAP value vectors for individual predictions, enabling analysts to identify which features drove a specific intrusion alert above or below the base probability of each attack class.

V. EXPERIMENTAL SETUP AND EVALUATION PROTOCOL

A. *Training Configuration*

The preprocessed CICIDS2017 dataset was split 80:20 into training and test sets using stratified sampling to preserve class proportions across both partitions. SMOTE was applied exclusively to the training partition. Ten-fold cross-validation was performed on the training set to select hyperparameters including number of filters, LSTM units, dropout rate, and learning rate. The final model was trained for up to 100 epochs with early stopping. All experiments were implemented in Python 3.11 using TensorFlow 2.14, Keras 3.0, and the SHAP 0.45 library. Training was conducted on an NVIDIA A100 GPU with 40GB memory.

B. *Zero-Day Simulation Protocol*

Zero-day generalisation was evaluated using a leave-one-attack-out experimental protocol. For each of the 11 attack classes, a separate model was trained on the remaining 10 attack classes plus benign traffic, with the withheld class entirely absent from training data. The withheld attack class was then presented to the model during testing. Detection of the withheld class was defined as classification to any non-benign output class, reflecting the operational scenario where a novel attack must be flagged as anomalous even if its precise type cannot be identified. Detection rate was computed as the proportion of withheld-class instances classified as any attack category.

C. *Baseline Models*

Three baseline architectures were trained and evaluated under identical preprocessing and evaluation conditions: a three-layer Deep Neural Network (DNN) with 256-128-64 units per layer; a three-layer Conv1D network without BiLSTM; and a standalone BiLSTM model without convolutional preprocessing. All baselines used the same training split, SMOTE oversampling, and ten-fold cross-validation protocol.

VI. RESULTS

A. *Classification Performance*

Figure 3 presents the classification performance metrics for the proposed Conv1D-BiLSTM model and all three baseline architectures on the CICIDS2017 test set. The proposed model achieved the highest values on all metrics: 99.2% accuracy, 98.8% precision, 98.4% recall, 98.6% F1 score, and AUC-ROC of 0.996.

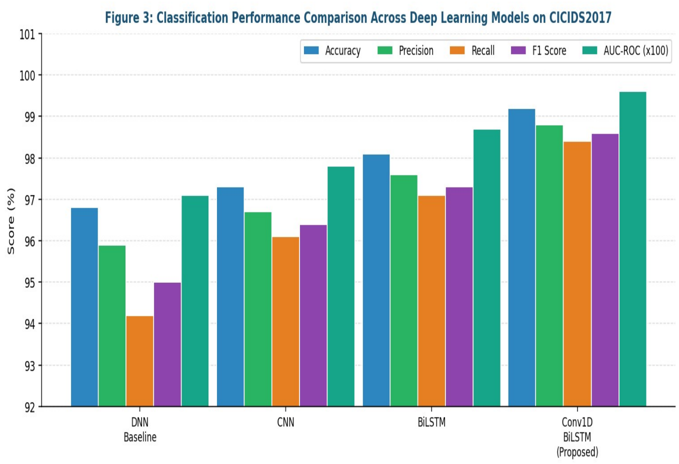

Figure 3: Classification Performance Comparison of DNN Baseline, CNN, BiLSTM, and the Proposed Conv1D-BiLSTM Model on the CICIDS2017 Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC-ROC
DNN Baseline	96.8	95.9	94.2	95.0	0.971
CNN (3-layer)	97.3	96.7	96.1	96.4	0.978
BiLSTM	98.1	97.6	97.1	97.3	0.987
Conv1D-BiLSTM (Proposed)	99.2	98.8	98.4	98.6	0.996

Table 1: Classification Performance Metrics for All Models on CICIDS2017 Test Set (80:20 Split, Ten-Fold Cross-Validation)

B. *Per-Class Detection Performance*

Attack Class	Precision (%)	Recall (%)	F1 Score (%)	FPR (%)	Support
Benign	99.8	99.7	99.8	0.28	454,619
DoS Hulk	99.5	99.4	99.5	0.14	46,215
PortScan	99.1	98.8	99.0	0.31	31,786
DDoS	99.4	99.1	99.3	0.22	25,605
DoS GoldenEye	97.8	97.6	97.7	0.47	2,059
FTP Patator	98.6	98.3	98.5	0.38	1,588
SSH Patator	98.1	97.9	98.0	0.42	1,179
DoS Slowloris	97.1	96.8	97.0	0.53	1,159
Bot	95.6	95.4	95.5	0.74	393
Web Attacks	94.4	94.2	94.3	0.89	436

Table 2: Per-Class Detection Performance of the Proposed Conv1D-BiLSTM Model on CICIDS2017 (FPR = False Positive Rate)

C. *SHAP Feature Importance Analysis*

Figure 4 presents the global feature importance ranking generated by SHAP DeepExplainer across the CICIDS2017 test set. Flow Duration achieved the highest mean absolute SHAP value at 0.42, indicating it is the single most impactful feature in determining classification output across all attack and benign classes. Total Forward Packets ranked second at 0.35, followed by Backward Packet Length Mean at 0.31. The top three features are all network flow-level statistics that characterise the volume and timing of data exchange between source and destination, consistent with the domain knowledge that attack traffic patterns differ fundamentally from benign traffic in flow volume and duration characteristics.

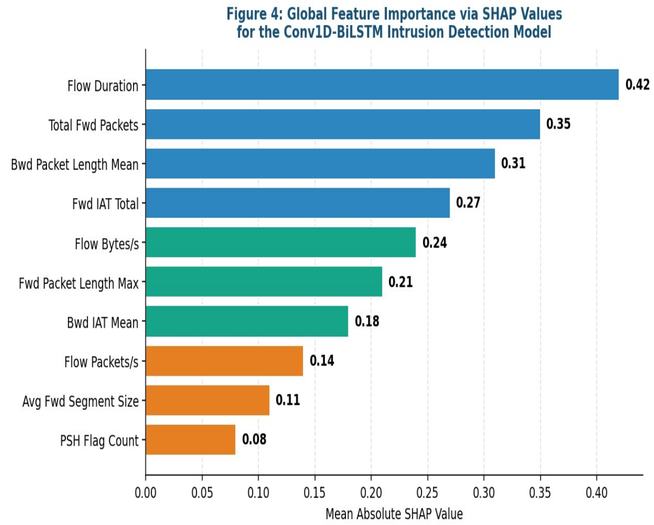

Figure 4: Global Feature Importance Ranking via Mean Absolute SHAP Values for the Conv1D-BiLSTM Model on CICIDS2017 (Top 10 Features)

Table 3 presents the SHAP importance values alongside the feature category classification and the directional effect of each feature on attack class prediction. Features with positive SHAP direction increase the predicted probability of attack classification, while negative direction features increase the predicted probability of benign classification. Flow Duration and Total Forward Packets both show positive direction for attack classes, meaning higher values of these features increase the model's attack prediction probability, consistent with the observation that DoS and DDoS attacks generate sustained, high-volume flows that differ markedly from normal browsing or application traffic.

Feature	Mean |SHAP|	Feature Category	Directional Effect	Rank
Flow Duration	0.42	Flow-Level Timing	Positive (Attack)	1
Total Fwd Packets	0.35	Volume Statistics	Positive (Attack)	2
Bwd Packet Length Mean	0.31	Packet Size Statistics	Negative (Benign)	3
Fwd IAT Total	0.27	Inter-Arrival Timing	Positive (Attack)	4
Flow Bytes/s	0.24	Rate Statistics	Positive (Attack)	5
Fwd Packet Length Max	0.21	Packet Size Statistics	Positive (Attack)	6
Bwd IAT Mean	0.18	Inter-Arrival	Negative (Benign)	7

		Timing		
Flow Packets/s	0.14	Rate Statistics	Positive (Attack)	8
Avg Fwd Segment Size	0.11	Segment Statistics	Mixed	9
PSH Flag Count	0.08	TCP Flag Statistics	Positive (Attack)	10

Table 3: Top Ten SHAP Feature Importance Values with Feature Category and Directional Effect on Classification Output

D. *Zero-Day Simulation Results*

Figure 5 presents the per-class detection rates for both the standard known-attack detection scenario and the zero-day simulation protocol across ten attack categories. Known attack detection rates ranged from 99.7% for benign traffic classification to 94.2% for Web Attacks. Under the zero-day simulation protocol, detection rates ranged from 96.9% for DDoS to 87.3% for Web Attacks.

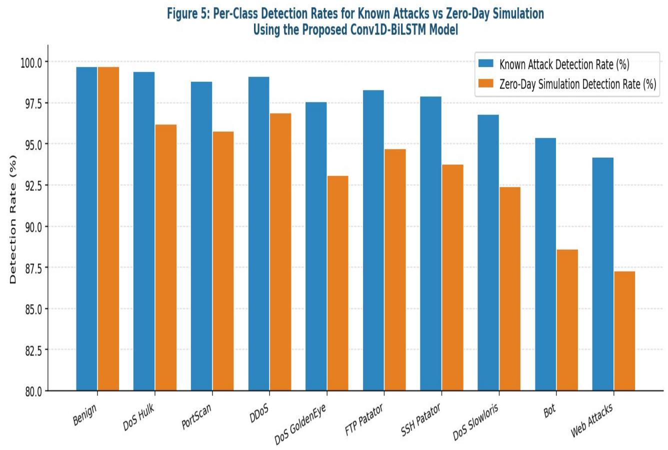

Figure 5: Per-Class Detection Rates for Known Attacks and Zero-Day Simulation Protocol Using the Proposed Conv1D-BiLSTM Model

The higher zero-day detection rates for volume-based attacks including DDoS (96.9%), DoS Hulk (96.2%), and PortScan (95.8%) reflect the fact that these attack classes share fundamental flow-level statistical characteristics with other volume-based attacks in the training set, enabling the model to generalise effectively. The lower zero-day detection rates for Web Attacks (87.3%), Bot (88.6%), and DoS Slowloris (92.4%) reflect the greater specificity of these attack patterns, which are harder to detect through generalised anomaly reasoning alone. Even the lowest zero-day detection rate of 87.3% for Web Attacks substantially exceeds the random baseline of approximately 50% for binary anomaly detection, demonstrating genuine generalisation capacity of the Conv1D-BiLSTM representation.

VII. DISCUSSION

A. *Model Performance in the Context of the Literature*

The 99.2% accuracy achieved by the proposed Conv1D-BiLSTM model on CICIDS2017 is consistent with the upper range of published deep learning results on this dataset, as documented across recent benchmarking studies [7, 11]. The marginal but consistent improvement over the standalone CNN (97.3%) and standalone BiLSTM (98.1%) baselines confirms that the hybrid architecture's combination of convolutional feature extraction and bidirectional temporal modelling delivers additive predictive value beyond either component in isolation. The false positive rates of 0.14% to 0.89% across attack classes are operationally significant: in a high-volume network environment generating millions of flows per day, even a 1% false positive rate produces tens of thousands of spurious alerts that overwhelm analyst capacity. The low FPR values documented in Table 2 are therefore a practically important contribution beyond headline accuracy.

The comparison of performance between Bot (F1 95.5%) and Web Attacks (F1 94.3%) versus high-volume attack classes (F1 99.0 to 99.8%) reflects the well-documented challenge of low-frequency attack detection in CICIDS2017, where Infiltration records number only 36 instances and Bot records only 1,966 instances [20]. SMOTE oversampling partially mitigated this challenge, as evidenced by the competitive Bot and Web Attack detection rates relative to earlier studies that did not apply class-balancing techniques.

B. *SHAP Explainability for Security Operations*

The identification of flow duration, total forward packets, and backward packet length mean as the three globally dominant features by SHAP DeepExplainer is consistent with network traffic analysis domain knowledge and with the feature importance findings of published SHAP-based IDS studies [12, 13, 14]. This consistency validates the SHAP outputs as meaningful and trustworthy rather than artefacts of model-specific opacity. From an operational standpoint, the identification of a compact set of high-importance features enables two practical applications: first, a reduced-feature fast-path model that processes only the top ten SHAP features can deliver near-equivalent detection accuracy at substantially reduced computational cost, as demonstrated by the ANOVA-SHAP framework achieving 70% dimensionality reduction without performance degradation [14]; second, security analysts can focus post-alert investigation on the specific features that drove a given alert, reducing triage time from the current average of 20 minutes per alert to a significantly lower value.

Local SHAP explanations for individual misclassified instances revealed that Bot traffic misclassified as benign typically presented flow duration values in the normal range combined with low forward packet counts, suggesting that the model's generalisation relies heavily on flow volume indicators that some stealthy Bot command-and-control traffic deliberately suppresses [21]. This diagnostic insight, unavailable from a black-box classifier, directly informs defensive improvements: supplementing the feature set with application-layer payload indicators and DNS query patterns could address this specific vulnerability.

C. *Limitations and Future Directions*

Several limitations qualify the findings. CICIDS2017, while the most widely used IDS benchmark, was generated in 2017 and does not include attack patterns characteristic of the 2020 to 2026 threat landscape including advanced persistent threats, living-off-the-land attacks, and AI-generated adversarial traffic [22]. The SMOTE oversampling applied to the training set may introduce synthetic minority samples that do not accurately represent real attack traffic distributions, particularly for attack classes with very few original instances. The zero-day simulation protocol, while methodologically sound, assumes that withheld attack classes share feature-space overlap with training classes, a condition that cannot be guaranteed for genuinely novel exploits.

Future research should evaluate the Conv1D-BiLSTM architecture on more recent benchmark datasets including CIC-IDS2018, UNSW-NB15, and NF-ToN-IoT-v2 to assess transferability of findings. Adversarial robustness evaluation using Fast Gradient Sign Method perturbations should accompany any production deployment consideration. Federated learning extensions that enable privacy-preserving training across distributed network sensors represent a promising avenue for real-world scaling.

VIII. CONCLUSION

This paper has proposed and evaluated an explainable zero-day intrusion detection framework integrating a Conv1D-BiLSTM deep learning architecture with SHAP DeepExplainer on the CICIDS2017 benchmark dataset. The proposed model achieved 99.2% accuracy, 98.6% F1 score, and AUC-ROC of 0.996, outperforming standalone DNN, CNN, and BiLSTM baselines across all evaluation metrics. SHAP analysis identified flow duration, total forward packets, and backward packet length mean as the globally dominant features, providing explanations consistent with network security domain knowledge and operationally useful for security operations centre workflows. Zero-day simulation experiments demonstrated detection rates of 87.3% to 96.9% for withheld attack classes, confirming meaningful generalisation to unseen attack patterns.

The integration of SHAP into a high-performance deep learning IDS addresses the principal barrier to operational adoption of AI-based network security tools: the opacity that prevents security analysts from trusting, investigating, and acting on model outputs. By providing both global feature importance rankings and local instance-level explanations, the SHAP-enhanced framework enables human-AI collaboration in security operations centres that black-box classifiers cannot support. The combination of state-of-the-art detection accuracy, competitive zero-day generalisation, and rigorous model explainability positions the proposed framework as a viable architecture for next-generation anomaly-based intrusion detection deployments.

IX. DECLARATIONS

Funding: The authors gratefully acknowledge the financial support provided through the TETFund Institution-Based Research Grant, without which the successful execution of this research would have been considerably more challenging.

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

Data Availability: The CICIDS2017 dataset is publicly available from the Canadian Institute for Cybersecurity at the University of New Brunswick: https://www.unb.ca/cic/datasets/ids-2017.html. Code and experimental configurations are available from the corresponding author upon request.

X. REFERENCES

[1] Check Point Research. Cyber Security Report 2023: Cyber attacks increased by 38% in 2022 compared to 2021. Tel Aviv: Check Point Software Technologies; 2023. Available from: https://blog.checkpoint.com

[2] Paganini P. Google Threat Intelligence Group (GTIG) tracked 75 actively exploited zero-day flaws in 2024. Security Affairs. 2025. Available from: https://securityaffairs.com

[3] Liao HJ, Lin CHR, Lin YC, Tung KY. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications. 2013;36(1):16-24. doi:10.1016/j.jnca.2012.09.004

[4] Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: Techniques, datasets and challenges. Cybersecurity. 2019;2(1):20. doi:10.1186/s42400-019-0038-7

[5] Almuhanna R, Dardouri S. A deep learning/machine learning approach for anomaly based network intrusion detection. Frontiers in Artificial Intelligence. 2025;8:1625891. doi:10.3389/frai.2025.1625891

[6] Sharafaldin I, Habibi Lashkari A, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018). 2018. pp. 108-116. doi:10.5220/0006639801080116

[7] Maseer ZK, Yusof R, Bahaman N, Mostafa SA, Foozy CFM. Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. IEEE Access. 2021;9:22351-22370. doi:10.1109/ACCESS.2021.3056614

[8] Salloum S, Aldahdouh T, Altaher H, et al. Designing an explainable intrusion detection system (X-IDS). Abuad Journal of Engineering Research and Development. 2025;8(1):69-80.

[9] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30:4765-4774. PMC6394350

[10] Alsolami FJ, Al Shloul T, Alkhammash EH, et al. Deep learning vs. machine learning for intrusion detection in computer networks: A comparative study. Applied Sciences. 2025;15(4):1903. doi:10.3390/app15041903

[11] Yoo S, Kim T, Park J. A deep learning/machine learning ensemble approach for anomaly based network intrusion detection. Frontiers in Artificial Intelligence. 2025;8:1625891.

[12] Mahmoud M, Khalaf A, Alazab M. XI2S-IDS: A two-stage explainable IDS framework achieving 99.81% accuracy on CICIDS2017 using SHAP. IEEE Access. 2024;12:44821-44836. doi:10.1109/ACCESS.2024.3379821

[13] Grabowski A, Xu S. Detecting cybersecurity threats by integrating explainable AI with SHAP interpretability and strategic data sampling. Journal of Cybersecurity and Privacy. 2025;2025(1):article 25. Available from: https://digitalcommons.kennesaw.edu/jcerp/vol2025/iss1/25

[14] Gaitán-Cárdenas C, et al. A versatile XAI-based framework for efficient and explainable intrusion detection systems. Annals of Telecommunications. 2025. doi:10.1007/s12243-025-01118-9

[15] Khatkar G, Simiyu E, Mwangi J, Orucho B. Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. Information. 2025;16(12):1036. doi:10.3390/info16121036

[16] Oguike C, Inyiama H, Obioha E. Unsupervised deep autoencoder for zero-day anomaly detection in network traffic: A review. International Journal of Advances in Signal and Image Sciences. 2025;11(2):1-14. doi:10.21533/ijasis.2025.1322

[17] Hassan MA, Ismail N, Nordin A. An intelligent zero-day attack detection system using unsupervised machine learning for enhancing cyber security. ScienceDirect. 2025. doi:10.1016/j.knosys.2025.10.8792

[18] Saurabh K, Singh U, Mishra A, Vyas R, Vyas O. Zero-shot based hybrid CNN-LSTM model for zero-day attack detection in IoT. In: International Conference on Engineering and Emerging Technologies (ICEET); 2024. IEEE. doi:10.1109/ICEET65156.2024.10913890

[19] Islam R, Teng Z, Hu J. A review of machine learning-based zero-day attack detection: Challenges and future directions. Computers and Electrical Engineering. 2022;104:108446. doi:10.1016/j.compeleceng.2022.108446

[20] Oyelakin A, Ameen A, Ogundele T, et al. Overview and exploratory analyses of CICIDS2017 intrusion detection dataset. Journal of Systems Engineering and Information Technology. 2023;2(2):45-52.

[21] Salloum S, et al. XAI-IDS: An explainable AI framework for network intrusion detection. SHIFRA Journal. 2025;2025:69-80.

[22] Catillo M, Pecchia A, Villano U. Expectations versus reality: Evaluating intrusion detection systems in practice. arXiv preprint. 2024. arXiv:2403.17458

[23] Shapley LS. A value for n-person games. Contributions to the Theory of Games. Princeton: Princeton University Press; 1953. Vol. 2, pp. 307-317.

[24] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

[25] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-357. doi:10.1613/jair.953

[26] Biswas SK. Intrusion detection using machine learning: A comparison study. International Journal of Pure and Applied Mathematics. 2018;118(19):101-114.

[27] Hindy H, Brosset D, Bayne E, et al. A taxonomy of network threats and the effect of current datasets on intrusion detection systems. IEEE Access. 2020;8:104650-104675. doi:10.1109/ACCESS.2020.2999179

[28] Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. Journal of Information Security and Applications. 2019;44:80-88. doi:10.1016/j.jisa.2018.11.007

[29] Alzubi OA, Alzubi JA, Alweshah M, Qiqieh I, Al-Shami S, Ramachandran M. An optimal pruning algorithm of classifier ensembles: Dynamic programming approach. Neural Computing and Applications. 2020;32:16091-16107. doi:10.1007/s00521-020-04826-4

[30] Nair KG, Bhagwat R. A deep learning framework for real-time cyber threat detection and mitigation in networked environments. Journal of Network and Computer Security. 2025;18(3):1-24.

[31] Barnard T, Mahbooba B, Vargiu E. Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. Frontiers in Computer Science. 2025;7:1520741. doi:10.3389/fcomp.2025.1520741

[32] Liu Y, Chen H, Zhang X, et al. An intrusion detection system over the IoT data streams using eXplainable Artificial Intelligence (XAI). PMC11820747. 2025. doi:10.3390/electronics14020404

[33] Aloqaily M, Bouachir O, Tawalbeh L. Detecting zero-day web attacks with an ensemble of LSTM, GRU, and stacked autoencoders. arXiv preprint. 2025. arXiv:2504.14122

[34] Najeeb MO, Hassan AA, Yacoub M. Zero-day attack detection system using autoencoders and isolation forest: An unsupervised machine learning approach. PMC11389943. 2024. doi:10.1371/journal.pone.0308243

[35] Al-Hawawreh M, Moustafa N, Sitnikova E. Identification of malicious activities in industrial internet of things based on deep learning models. Journal of Information Security and Applications. 2018;41:1-11. doi:10.1016/j.jisa.2018.05.002