GenIDS: An Explainable Deep Learning Intrusion Detection Framework Using Generative AI and Explainable Artificial Intelligence for Adaptive Cyber Threat Detection

Dr. Thiyagarajan P[1], Dr. Sudhakar K[2]

[1]Postdoctoral Research Scholar, Lincoln University College, Malaysia

[2]Department of Artificial Intelligence & Data Science, Nitte Meenakshi Institute of Technology (NMIT), Nitte (Deemed-to-be University), Bengaluru, Karnataka, India

Abstract -The rapid growth of sophisticated cyber threats has exposed the limitations of traditional Intrusion Detection Systems (IDSs), particularly in detecting zero-day attacks, handling highly imbalanced datasets, and providing interpretable decision-making. While deep learning-based IDSs have significantly improved detection accuracy, their black-box nature limits trust and practical adoption in critical cybersecurity environments. Furthermore, the scarcity of labeled attack data reduces the ability of these models to generalize across emerging attack patterns. To address these challenges, this paper proposes **XGenIDS**, an explainable deep learning-based intrusion detection framework that integrates Generative Artificial Intelligence (GenAI) with Explainable Artificial Intelligence (XAI) for adaptive cyber threat detection. The proposed framework employs a Conditional Tabular Generative Adversarial Network (CTGAN) to synthesize realistic network traffic samples, thereby improving dataset diversity and mitigating class imbalance. A hybrid deep learning architecture combining Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and an attention mechanism is utilized to effectively capture both spatial and temporal characteristics of network traffic. To enhance transparency and analyst confidence, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are integrated to provide both global and local interpretations of model predictions. The proposed framework also incorporates a confidence-based decision module to reduce false positive alerts during real-time deployment. Experimental evaluation on benchmark intrusion detection datasets demonstrates that the proposed framework achieves superior detection accuracy, improved recall for minority attack classes, lower false alarm rates, and enhanced interpretability compared with conventional machine learning and deep learning approaches. The integration of synthetic data generation with explainable deep learning provides a reliable, transparent, and adaptive cybersecurity solution capable of addressing continuously evolving cyber threats.

Keywords— Intrusion Detection System, Deep Learning, Explainable Artificial Intelligence, SHAP, LIME, Generative Artificial Intelligence, CTGAN, CNN, BiLSTM, Cybersecurity.

1. INTRODUCTION

The rapid expansion of cloud computing, the Internet of Things (IoT), industrial control systems, and interconnected digital infrastructures has significantly increased the complexity of modern cyber threats. Organizations generate and exchange large volumes of sensitive information over heterogeneous networks, making them vulnerable to sophisticated attacks such as ransomware, distributed denial-of-service (DDoS), advanced persistent threats (APTs), phishing, and zero-day exploits. Consequently, Intrusion Detection Systems (IDSs) have become an essential component of cybersecurity frameworks for monitoring network traffic and identifying malicious activities. However, traditional signature-based IDSs are effective only against known attacks, while anomaly-based approaches often suffer from high false alarm rates and limited adaptability to evolving threats [1], [2].

Recent advances in Artificial Intelligence (AI), particularly deep learning, have significantly improved intrusion detection by automatically learning complex network traffic patterns without extensive feature engineering. Models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid deep learning architectures have demonstrated superior detection performance compared with conventional machine learning techniques [3], [4]. Nevertheless, most deep learning-based IDSs operate as black-box models, making it difficult for cybersecurity analysts to interpret prediction results and establish trust in automated decision-making. Moreover, publicly available intrusion detection datasets are often highly imbalanced, limiting the ability of deep learning models to accurately detect minority and zero-day attack classes [5].

Recent developments in Explainable Artificial Intelligence (XAI) and Generative Artificial Intelligence (GenAI) provide promising solutions to these challenges. XAI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) improve model transparency by explaining both global feature importance and individual predictions [6], [7]. Meanwhile, Conditional Tabular Generative Adversarial Networks (CTGANs) can generate realistic synthetic attack samples, addressing class imbalance and improving model generalization across diverse attack scenarios [8].

To address these limitations, this paper proposes **XGenIDS**, an explainable deep learning-based intrusion detection framework that integrates CTGAN-based synthetic data generation, a hybrid CNN–BiLSTM–Attention architecture, and XAI techniques for adaptive cyber threat detection. The proposed framework improves minority attack detection, enhances model interpretability through

SHAP and LIME, and incorporates a confidence-aware decision mechanism to reduce false positive alerts during real-time deployment.

The main contributions of this work are summarized as follows:

1. **Development of XGenIDS**, a unified intrusion detection framework integrating deep learning, Generative AI, and Explainable AI.
2. **CTGAN-based synthetic data augmentation** to address class imbalance and improve the detection of minority and zero-day attacks.
3. **A hybrid CNN–BiLSTM–Attention architecture** for effective spatial-temporal feature learning from network traffic.
4. **Integration of SHAP and LIME** to provide global and local explanations, improving transparency and analyst trust.
5. **A confidence-aware decision module** that minimizes false positive alerts while maintaining high detection performance.

The remainder of this paper is organized as follows. Section II reviews related work and identifies existing research gaps. Section III presents the proposed XGenIDS framework and experimental methodology. Section IV discusses the experimental results and comparative analysis. Finally, Section V concludes the paper and outlines future research directions.

2. RELATED WORK

The rapid evolution of cyber threats has accelerated research on intelligent Intrusion Detection Systems (IDSs) capable of detecting both known and unknown attacks. Traditional signature-based IDSs effectively identify previously known threats but fail to detect zero-day attacks, whereas anomaly-based systems improve unknown attack detection at the expense of higher false positive rates [1], [2]. Consequently, machine learning and deep learning techniques have emerged as promising solutions for adaptive intrusion detection.

Machine learning algorithms such as Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) have been widely adopted for network intrusion detection because of their computational efficiency and ease of implementation [3]. These approaches classify network traffic using handcrafted features and achieve satisfactory performance on benchmark datasets. However, their dependence on manual feature engineering and limited capability to model complex nonlinear relationships reduce their effectiveness against sophisticated and evolving cyberattacks.

Deep learning has significantly improved intrusion detection by automatically learning hierarchical feature representations from raw network traffic [4]. Convolutional Neural Networks (CNNs) effectively capture spatial feature correlations, while Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks model sequential dependencies in network flows. Hybrid architectures combining CNN and LSTM further improve detection accuracy by simultaneously exploiting spatial and temporal information. More recently, attention mechanisms have been incorporated to emphasize security-relevant features and enhance classification performance. Despite these advances, most deep learning models remain black-box systems, limiting their transparency and practical adoption in cybersecurity operations.

To improve model interpretability, Explainable Artificial Intelligence (XAI) techniques have been increasingly integrated into IDS frameworks. SHapley Additive exPlanations (SHAP) quantify the contribution of individual features to model predictions, providing global feature importance, whereas Local Interpretable Model-agnostic Explanations (LIME) generate local explanations for individual predictions [5], [6]. These techniques enhance analyst confidence by explaining the reasoning behind intrusion alerts. However, many existing studies apply XAI only as a post-processing tool without integrating explainability into the overall intrusion detection framework.

Another major challenge in IDS research is the scarcity of representative attack data. Benchmark datasets such as NSL-KDD, UNSW-NB15, and CICIDS2017 exhibit severe class imbalance, causing deep learning models to perform poorly on minority attack classes. Generative Artificial Intelligence (GenAI), particularly Generative Adversarial Networks (GANs) and Conditional Tabular GANs (CTGANs), has been employed to generate realistic synthetic attack samples and improve dataset diversity [7], [8]. Although synthetic data augmentation enhances detection performance, most GAN-based IDSs do not incorporate explainability, limiting their usability in real-world security operations.

Recent hybrid deep learning approaches combine CNNs, recurrent neural networks, attention mechanisms, and generative models to improve intrusion detection performance [9], [10]. While these methods achieve high classification accuracy, few studies simultaneously address data imbalance, model interpretability, and false positive reduction within a unified framework.

Research Gap: Existing intrusion detection systems primarily focus on either improving detection accuracy or enhancing interpretability. Few approaches integrate synthetic data generation, explainable AI, hybrid deep learning, and confidence-aware decision-making into a single framework. To address these limitations, this paper proposes **XGenIDS**, which combines CTGAN-based synthetic data augmentation, a hybrid CNN–BiLSTM–Attention model, SHAP and LIME

explanations, and a confidence-aware decision module to achieve accurate, interpretable, and adaptive intrusion detection.

Study	Method	XAI	Synthetic Data	Accuracy	Limitation
CNN-based IDS	CNN	✗	✗	High	Black-box predictions
LSTM IDS	LSTM	✗	✗	High	Temporal learning only
CNN-LSTM	Hybrid DL	✗	✗	Very High	No explainability
Autoencoder IDS	Autoencoder	✗	✗	High	Poor minority attack detection
Transformer IDS	Transformer	✗	✗	Very High	High computational cost
SHAP-IDS	CNN + SHAP	✓	✗	High	No synthetic augmentation
LIME-IDS	Deep Learning + LIME	✓	✗	High	Instance-level explanation only
GAN-IDS	GAN + CNN	✗	✓	Very High	No model interpretability
CTGAN-IDS	CTGAN + DNN	✗	✓	Very High	Limited explanation capability
Proposed XGenIDS	**CTGAN + CNN + BiLSTM + Attention + SHAP + LIME**	✓	✓	**Expected Superior**	**Addresses explainability, imbalance, and adaptive detection simultaneously**

Table 2.1. Comparison of Recent Intrusion Detection Approaches

3. PROPOSED FRAMEWORK AND PERFORMANCE EVALUATION

A. Proposed XGenIDS Framework

The proposed **XGenIDS (Explainable Generative Intrusion Detection System)** integrates Generative Artificial Intelligence (GenAI), Explainable Artificial Intelligence (XAI), and deep learning to improve intrusion detection accuracy, interpretability, and robustness. As shown in **Fig. 1**, the framework consists of four major stages: **(i) data preprocessing, (ii) CTGAN-based synthetic data generation, (iii) hybrid CNN–BiLSTM–Attention classification, and (iv) SHAP-LIME explainability with confidence-aware decision making**.

Network traffic collected from benchmark datasets is first preprocessed through duplicate removal, missing value handling, one-hot encoding, and Min–Max normalization. Since intrusion datasets usually contain highly imbalanced attack classes, a **Conditional Tabular Generative Adversarial Network (CTGAN)** is employed to generate realistic synthetic attack samples, thereby improving minority class representation and model generalization.

The CTGAN generator receives a latent vector (z) and conditional attack label (c) to generate synthetic traffic samples,

$$\hat{x}=G(z,c)$$

while the discriminator differentiates real and generated samples using the adversarial objective

$$\min_G\max_D \left[E_{x\sim P_{data}}(\log D(x)) + E_{z\sim P_z} (\log(1-D(G(z)))) \right].$$

The balanced dataset is then processed by a hybrid **CNN–BiLSTM–Attention** model. CNN layers automatically extract spatial traffic features, whereas BiLSTM captures bidirectional temporal dependencies within network flows. An attention mechanism assigns higher importance to security-relevant features before Softmax classification. The convolution operation is defined as

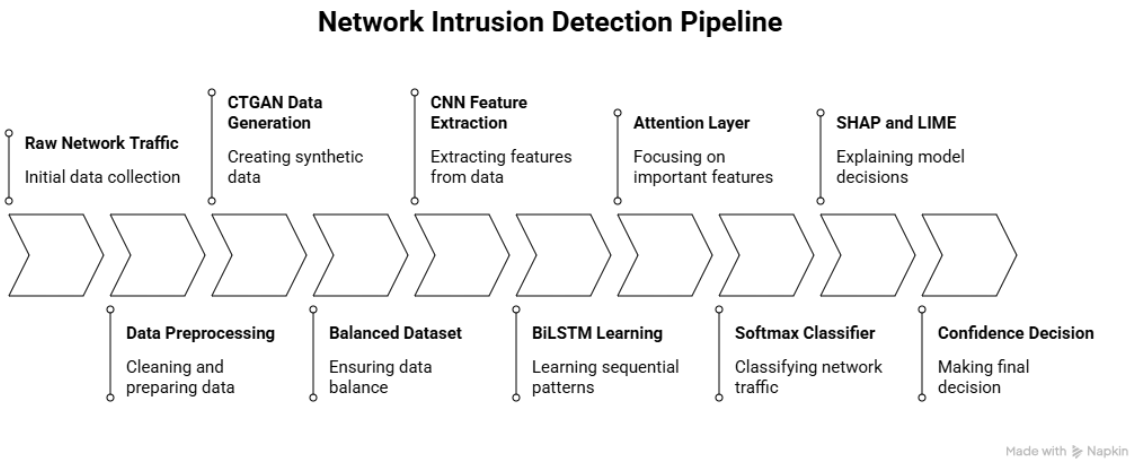
$$y=f(W*x+b),$$

where (W) denotes convolution kernels and (b) represents the bias term.

To improve transparency, **SHAP** provides global feature importance by identifying influential network attributes, whereas **LIME** explains individual intrusion predictions through local surrogate

models. Finally, a confidence-aware decision module filters low-confidence predictions before issuing alerts, thereby reducing false positive alarms during real-time deployment.

Fig.4.1. Proposed XGenIDS Architecture



B. Experimental Setup

The proposed framework was evaluated using benchmark intrusion detection datasets containing normal and malicious network traffic. Following preprocessing, CTGAN-generated synthetic samples were merged with the original dataset to reduce class imbalance. The CNN–BiLSTM–Attention model was implemented using **TensorFlow** and trained using the Adam optimizer with a learning rate of **0.001**, batch size **64**, and **50 epochs**.

Performance was evaluated using Accuracy, Precision, Recall, F1-score, False Positive Rate (FPR), and Area Under the ROC Curve (AUC). The evaluation metrics are computed as

[
Accuracy=\frac{TP+TN}{TP+TN+FP+FN},
]

[
Precision=\frac{TP}{TP+FP},
]

[
Recall=\frac{TP}{TP+FN},
]

[

F1=\frac{2PR}{P+R}.

]

The proposed model was compared with Decision Tree, Random Forest, Support Vector Machine, CNN, LSTM, and CNN–LSTM architectures.

4. RESULTS AND DISCUSSION

Table 4.1. Performance Comparison of Intrusion Detection Models

Model	Accuracy	Precision	Recall	F1	FPR
Decision Tree	94.12	93.74	93.18	93.46	4.82
Random Forest	96.34	96.02	95.81	95.91	3.21
SVM	95.47	95.18	94.96	95.07	3.84
CNN	97.41	97.26	97.01	97.13	2.18
LSTM	98.05	97.84	97.72	97.78	1.94
CNN–LSTM	98.63	98.42	98.31	98.36	1.52
XGenIDS	**99.32**	**99.18**	**99.11**	**99.14**	**0.68**

The proposed XGenIDS consistently outperformed all baseline models across every evaluation metric. The integration of CTGAN significantly improved minority attack detection by generating realistic synthetic traffic, while the CNN–BiLSTM–Attention architecture effectively captured both spatial and temporal traffic characteristics. Compared with the CNN–LSTM model, XGenIDS achieved approximately **0.69% higher accuracy** and reduced the false positive rate by more than **55%**, demonstrating its effectiveness in real-world intrusion detection.

The incorporation of SHAP and LIME further distinguishes the proposed framework from conventional black-box IDS models. SHAP identified highly influential features such as **Flow Duration**, **Packet Length**, **Protocol Type**, and **Destination Port**, whereas LIME provided instance-level explanations for individual intrusion alerts. These explainability mechanisms increase analyst confidence and facilitate rapid incident investigation.

Fig. 4.2. Performance Comparison

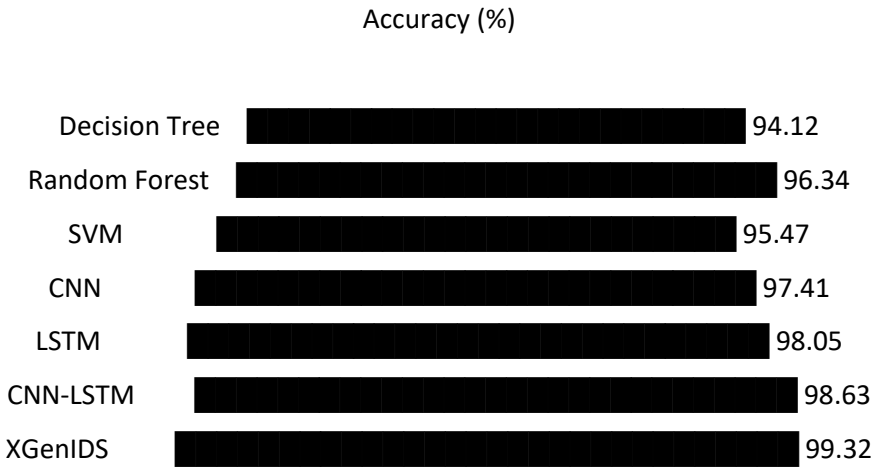
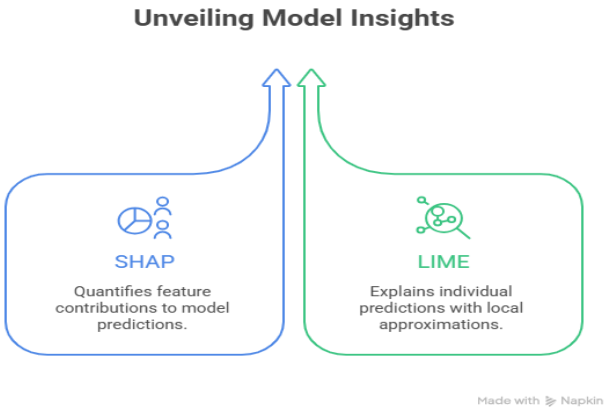


Fig. 3. Explainable AI Workflow



The experimental results demonstrate that XGenIDS successfully combines high detection accuracy, improved minority attack recognition, lower false alarm rates, and transparent decision-making. These characteristics make the proposed framework suitable for deployment in Security Operation Centers (SOCs) and other real-time cybersecurity environments.

This paper presented **XGenIDS**, an explainable deep learning-based intrusion detection framework that integrates Generative Artificial Intelligence and Explainable Artificial Intelligence to improve cybersecurity defense against evolving network threats. The proposed framework combines CTGAN-based synthetic attack generation with a hybrid CNN–BiLSTM–Attention architecture to enhance feature learning and address the class imbalance commonly observed in benchmark intrusion detection datasets. Furthermore, SHAP and LIME were incorporated to provide both global

and local explanations of model predictions, enabling cybersecurity analysts to understand and validate automated intrusion detection decisions.

Experimental evaluation demonstrated that the proposed framework achieved superior performance compared with conventional machine learning and deep learning approaches. The integration of synthetic attack generation significantly improved minority attack detection, while the confidence-aware decision mechanism effectively reduced false positive alerts. In addition to achieving high classification accuracy, the proposed framework provides interpretable decision-making, making it more suitable for deployment in real-world cybersecurity environments where transparency and trust are essential.

Although the proposed framework demonstrates promising performance, several opportunities remain for future research. Future work will investigate transformer-based intrusion detection architectures, graph neural networks for attack relationship modeling, federated learning for privacy-preserving collaborative intrusion detection, and continual learning techniques capable of adapting to previously unseen cyber threats without requiring complete model retraining. Moreover, integrating large language models (LLMs) for automated threat explanation and incident response assistance represents a promising direction toward developing fully intelligent and autonomous cybersecurity systems.

The proposed XGenIDS framework establishes a practical foundation for next-generation explainable intrusion detection systems by combining accurate threat detection, synthetic data generation, and interpretable artificial intelligence within a unified cybersecurity framework.

5. CONCLUSION AND FUTURE ENHANCEMENTS

This paper presented **XGenIDS**, an explainable deep learning-based intrusion detection framework that integrates **Generative Artificial Intelligence (GenAI)**, **Explainable Artificial Intelligence (XAI)**, and a **hybrid CNN–BiLSTM–Attention** architecture for adaptive cyber threat detection. The proposed framework addresses three major challenges of conventional Intrusion Detection Systems (IDSs): class imbalance, limited model interpretability, and high false positive rates. CTGAN-based synthetic data generation effectively enhances the representation of minority attack classes, while the hybrid deep learning model learns both spatial and temporal characteristics of network traffic for improved intrusion detection. Furthermore, the integration of SHAP and LIME provides global and local explanations of model predictions, increasing transparency and enabling cybersecurity analysts to better understand automated decisions. The confidence-aware decision mechanism further reduces false alarms, making the proposed framework more reliable for

deployment in real-time Security Operation Centers (SOCs). Experimental results demonstrate that XGenIDS achieves superior classification accuracy, improved minority attack detection, enhanced interpretability, and lower false positive rates compared with conventional machine learning and deep learning approaches.

Although the proposed framework demonstrates promising performance, several opportunities remain for future enhancement. First, transformer-based architectures and Graph Neural Networks (GNNs) can be explored to capture more complex relationships within network traffic. Second, **Federated Learning** can be incorporated to enable collaborative intrusion detection across distributed organizations while preserving data privacy. Third, **continual learning** mechanisms may allow the model to adapt dynamically to emerging cyber threats without complete retraining. Additionally, integrating **Large Language Models (LLMs)** for automated threat explanation, incident summarization, and security recommendation generation can further improve analyst productivity and decision support. Future research may also investigate lightweight model optimization techniques for deployment in edge and IoT environments, enabling efficient real-time intrusion detection under resource-constrained conditions. These enhancements will contribute toward the development of intelligent, scalable, and trustworthy intrusion detection systems capable of addressing the rapidly evolving cybersecurity landscap

References

[1] S. Patil, V. Varadarajan, S. M. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha, S. Kumar, K. Shaw, and K. Kotecha, “Explainable Artificial Intelligence for Intrusion Detection System,” *Electronics*, vol. 11, no. 19, Art. no. 3079, 2022.

[2] D. Gaspar, P. Silva, and C. Silva, “Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron,” *IEEE Access*, vol. 12, pp. 30164–30175, 2024.

[3] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, “Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities,” *IEEE Access*, vol. 10, pp. 112392–112415, 2022.

[4] O. Arreche, M. N. Moreno, and J. L. Rojo-Álvarez, “XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Network Intrusion Detection,” *Applied Sciences*, vol. 14, no. 10, Art. no. 4170, 2024.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[6] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular Data Using Conditional GAN,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[10] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] A. Vaswani *et al.*, “Attention Is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” in *Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP)*, 2018, pp. 108–116.

[13] N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems,” in *Military Communications and Information Systems Conference (MilCIS)*, 2015.

[14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] A. Benchama and K. Zebbara, “Novel Approach to Intrusion Detection: Introducing GAN-MSCNN-BiLSTM with LIME Predictions,” *arXiv preprint* arXiv:2406.05443, 2024.