MSI Journal of AI and Technology

Volume: 2, Issue: 3 (July-September 2026)
Published by *MSI Publishers*
ISSN 3107-6181 (Online)

Received: 25-07-2026
Accepted: 02-09-2026
Published: 05-09-2026

Integrating eXplainable Artificial Intelligence (XAI) and Deep Learning in Real-Time Network Intrusion Detection Systems: A Multi-Context Feature Shift Evaluation

Authors Details

Luong Truong An[1*]
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Vu Thanh Nhan[2]
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Corresponding Author: *Luong Truong An*

ABSTRACT: The exponential growth of network data and sophisticated cyber threats requires Intrusion Detection Systems (IDS) to possess both high accuracy and clear interpretability. This research proposes a comprehensive IDS framework that combines a hybrid Deep Learning architecture (CNN-LSTM) with the Synthetic Minority Over-sampling Technique (SMOTE) to efficiently identify anomalous behaviors. The focal point of this paper is the integration of the SHapley Additive exPlanations (SHAP) tool and adversarial evaluation approaches to break the "black-box" barrier of Deep Learning, providing transparent interpretations for every system decision. The system is empirically evaluated on modern benchmark datasets, demonstrating high performance in maintaining accuracy, enhancing human trust management, and tracking multi-context feature shifts.

Keywords: *Intrusion Detection Systems (IDS), Explainable Artificial Intelligence (XAI), Deep Learning, SHAP, SMOTE, Concept Drift, Trust Management*

1. Introduction

The rapid and ubiquitous expansion of digital connectivity, cloud computing architectures, and the Internet of Things (IoT) has fundamentally transformed global communication,

DOI: *10.5281/zenodo.22333206*

facilitating the exchange of petabytes of data on a daily basis [11, 20]. However, this hyper-connectivity has simultaneously expanded the cyber threat landscape, exposing critical national infrastructures and enterprise environments to increasingly sophisticated, automated, and evasive cyberattacks [10, 20]. In response, Intrusion Detection Systems (IDS) have evolved as an indispensable second line of defense, tasked with automatically monitoring, detecting, and categorizing anomalous or malicious behaviors before systemic damage occurs [10, 11].

Traditionally, signature-based IDSs struggle against zero-day exploits and polymorphic malware, driving the academic and industrial communities toward data-driven, machine-learning-based intrusion detection [8, 10]. More recently, deep learning (DL) models-such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) including Long Short-Term Memory (LSTM) networks-have achieved state-of-the-art performance by automatically extracting hierarchical feature representations from raw network traffic without relying on manual feature engineering [8, 10].

Despite their exceptional predictive accuracy, deploying deep learning models within operational Security Operations Centers (SOCs) remains constrained by two major bottlenecks:

1. **Data Imbalance and Generalization:** Real-world network traffic is inherently skewed, where benign traffic vastly overwhelms underrepresented minority attack classes [1, 8]. Standard classifiers trained on such data naturally suffer from high false-negative rates on rare intrusions [1]. While preprocessing methods like the Synthetic Minority Over-sampling Technique (SMOTE) help rebalance training distributions, they introduce complex density alterations in the feature space [1, 6].

2. **The "Black-Box" Dilemma and Trust Deficit:** Deep neural networks operate as opaque black boxes, providing final classification outputs without contextual reasoning or causal justifications [12]. In high-stakes security domains, human administrators require transparent explanations before executing automated mitigation policies. Without interpretability, high accuracy alone is insufficient to secure human trust and compliance [12, 13].

 DOI: *10.5281/zenodo.22333206*

Furthermore, operational networks are non-stationary environments where data distributions shift over time-a phenomenon known as *concept drift* or *covariate shift*-which rapidly degrades the performance of static machine learning models [2].

To bridge the gap between predictive accuracy and operational transparency, this research proposes a comprehensive, explainable intrusion detection framework. The primary contributions of this work are summarized as follows:

A Hybrid CNN-LSTM Architecture with SMOTE Preprocessing: We implement a robust sequential-spatial deep learning model augmented by SMOTE to effectively handle severe class imbalances and capture long-range temporal dependencies within network flow records [1, 6, 8, 10].

Integration of Explainable AI (XAI) via SHAP: We incorporate SHapley Additive exPlanations (SHAP) to unpack the black-box nature of the deep neural network, providing rigorous local (per-instance) and global (feature-importance) interpretations rooted in cooperative game theory [7, 9].

Adversarial Diagnostic Evaluation: We introduce an adversarial evaluation technique to probe the decision boundaries of the model, exposing the underlying feature deviations responsible for misclassifications and false alarms [12].

Multi-Context Feature Shift Analysis: We perform cross-dataset evaluation across benchmark environments (CICIDS2017 and UNSW-NB15) and provide **visual comparative XAI analyses** to track how feature importance dynamically redistributes under multi-context distributional shifts.

2. Related work

The literature surrounding intrusion detection and machine learning is extensive, yet it exhibits a distinct methodological dichotomy between pure performance optimization and model interpretability. This section systematically reviews existing advancements across three critical dimensions: deep-learning-based IDSs, data imbalance mitigation, and Explainable AI (XAI) in cybersecurity.

 DOI: *10.5281/zenodo.22333206*

2.1 Deep Learning Architectures for Intrusion Detection

Early machine learning applications in network security relied on shallow models, including Decision Trees, Support Vector Machines (SVM), and Naive Bayes classifiers [9, 20]. While computationally light, these models struggle to process high-dimensional modern traffic streams and fail to capture sequential dependencies across packet headers. To overcome these limitations, deep learning paradigms have gained dominance [10]. Vinayakumar et al. [8] demonstrated that deep neural networks (DNNs) significantly outperform classical classifiers across multiple benchmark datasets by leveraging multi-layer feature abstraction.

More recently, hybrid architectures that merge spatial feature extractors with temporal sequence models have emerged. For instance, convolutional layers are frequently coupled with LSTM or Gated Recurrent Units (GRU) to analyze spatial network interactions alongside temporal packet arrival patterns [8, 10]. Despite achieving high detection rates, these studies predominantly treat network security as a closed-form classification problem, largely overlooking the operational necessity of model interpretability and resilience against distribution shifts [8].

2.2 Mitigating Class Imbalance in Cyber Traffic

Class imbalance remains a foundational obstacle in anomaly-based intrusion detection, where minority attacks constitute a minute fraction of overall network traffic [1, 8]. Standard resampling strategies, such as random oversampling by replication, frequently lead to severe overfitting by restricting decision boundaries to specific data points [1, 6].

To address this, Chawla et al. [6] introduced the Synthetic Minority Over-sampling Technique (SMOTE), which constructs synthetic instances via linear interpolation within the feature space [1, 6]. Over the past two decades, numerous extensions of SMOTE have been proposed-such as Borderline-SMOTE, ADASYN, and MWMOTE-aimed at refining instance selection around class boundaries and mitigating noise [1]. While SMOTE and its variants successfully improve the sensitivity of classifiers toward underrepresented attacks, their interaction with complex deep learning architectures and high-dimensional flow representations requires careful regulation to avoid introducing synthetic noise into overlapping class regions [1, 6].

 DOI: *10.5281/zenodo.22333206*

2.3 Explainable Artificial Intelligence (XAI) and Trust Management in IDS

As AI systems assume greater autonomy in critical infrastructure protection, the demand for transparent and trustworthy models has surged [12, 13]. Traditional feature importance metrics offer global insights but fail to explain individual, instance-level predictions made by complex neural networks [9].

To establish human-AI collaboration and robust trust management, recent works have begun exploring post-hoc interpretability frameworks [12, 13]. Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) have surfaced as prominent tools for approximating black-box models locally [7, 9]. Lundberg and Lee [7] established a unified mathematical framework proving that SHAP values are the unique additive feature attribution measures satisfying local accuracy, missingness, and consistency based on cooperative game theory.

Within the cybersecurity domain, Wang et al. [9] and Marino et al. [12] demonstrated the utility of explainable frameworks for debugging classifier errors and understanding feature contributions. However, existing literature lacks a comprehensive evaluation that bridges deep-learning-based intrusion detection, rigorous XAI-driven feature validation, and the tracking of multi-context feature shifts across heterogeneous network environments [4, 5, 8]. This research directly addresses this gap.

3. Proposed Methodology and System Architecture

The proposed framework is structured into a rigorous end-to-end pipeline designed to address class imbalance, extract complex spatio-temporal representations from network traffic, and deliver model transparency through Explainable AI (XAI) [1, 6, 7, 8].

3.1 Data Preprocessing and Imbalance Mitigation

To prepare raw packet captures (PCAPs) for deep learning classification, network flows are processed using robust feature extraction tools such as CICFlowMeter [4]. Let the preprocessed dataset be denoted as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^F$ represents the vector of F network flow attributes and y_i corresponds to the traffic class label [4, 5].

 DOI: *10.5281/zenodo.22333206*

Because network intrusion datasets exhibit severe skewness, standard classifiers naturally bias toward the majority class [1]. To counteract this, SMOTE is applied exclusively to the training partition [1, 6]. SMOTE operates in the feature space by identifying the k-nearest neighbors for each minority instance $\mathbf{x}_i$ and synthesizing new instances $\mathbf{x}_{new}$ via random linear interpolation:

$$\mathbf{x}_{new} = \mathbf{x}_i + \delta(\mathbf{x}_z - \mathbf{x}_i) \quad (1)$$

where $\mathbf{x}_z$ is a randomly selected neighbor from the k-nearest neighbors of $\mathbf{x}_i$, and $\delta \in [0,1]$ is a random scalar [1, 6].

3.2 Hybrid CNN-LSTM Deep Learning Architecture

To capture both spatial local correlations and long-range temporal dependencies within network traffic flows, a hybrid 1D-CNN and LSTM architecture is implemented [8, 10].

Spatial Feature Extraction (1D-CNN): The normalized feature vectors are passed through one-dimensional convolutional layers to extract localized interaction patterns [8, 10]:

$$\mathbf{h}_t^{(conv)} = \mathrm{ReLU}(\mathbf{W}_c * \mathbf{x}_t + \mathbf{b}_c) \quad (2)$$

Temporal Dependency Modeling (LSTM): Stacked LSTM layers retain historical packet sequence information over extended temporal windows, making the model robust against prolonged intrusions [8, 10].

3.3 Explainability Module via SHAP

To resolve the opaque "black-box" nature of the hybrid deep neural network, the SHAP framework is integrated [7, 9]. Grounded in cooperative game theory, SHAP computes the marginal contribution of each feature [7]:

$$g(\mathbf{z}') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j' \quad (3)$$

where $\phi_j \in \mathbb{R}$ represents the SHAP value (marginal impact) of the j-th feature [7]

 DOI: *10.5281/zenodo.22333206*

3.4 Adversarial Diagnostic Framework

To audit decision boundaries and explain misclassifications, an adversarial perturbation approach is incorporated [12]. For a misclassified input sample $\mathbf{x}_0$, the framework computes the minimal input modification $\hat{\mathbf{x}}$ required to correct the prediction [12]:

$$\min_{\hat{\mathbf{x}}} \mathcal{L}\big(\hat{y}, f(\hat{\mathbf{x}})\big) + \lambda(\hat{\mathbf{x}} - \mathbf{x}_0)^T \mathbf{Q}(\hat{\mathbf{x}} - \mathbf{x}_0)\big) \quad (4)$$

The deviation vector $(\mathbf{x}_0 - \hat{\mathbf{x}})$ highlights feature vulnerabilities to assist engineers in hardening the IDS [12].

4. Experimental Setup

The proposed framework was implemented using Python (TensorFlow/Keras backend and Scikit-Learn) and evaluated on a Google Colab Pro environment equipped with GPU acceleration [8].

4.1 Datasets and Evaluation Metrics

1. **CICIDS2017:** Contains up-to-date network flow traffic with over 80 extracted features, encompassing benign traffic and diverse attack categories [4].
2. **UNSW-NB15:** Utilized for multi-context validation and robustness testing. Unlike CICIDS2017, this dataset incorporates distinct network topologies and connection state features, making it an ideal benchmark for evaluating cross-dataset feature shifts [5].

Model performance was evaluated using standard multi-class classification metrics: Accuracy, Precision, Recall, F1-Score, and Receiver Operating Characteristic - Area Under Curve (ROC-AUC) [3, 4, 8].

5. Results and Discussion

5.1 Classification Performance Evaluation

The quantitative classification performance of the proposed hybrid CNN-LSTM framework evaluated on the CICIDS2017 test set is comprehensively detailed in Table I.

DOI: *10.5281/zenodo.22333206*

The model achieves an overall classification accuracy of **95.89%** alongside a macro-averaged F1-score of **0.697** and a weighted-averaged F1-score of **0.966**.

Table 1. Classification Performance Metrics of the Hybrid CNN-LSTM Model on the CICIDS2017 Dataset

Class/Attack Category	Precision	Recall	F1-Score	Support
BENIGN	1	0.9	0.974	22671
Bot	0.06	0.952	0.113	21
DDoS	0.93	0.998	0.963	1269
DoS GoldenEye	0.687	1	0.815	101
DoS Hulk	0.943	0.999	0.97	2341
DoS Slowhttptest	0.775	0.965	0.859	57
DoS slowloris	0.824	1	0.903	56
FTP-Patator	0.911	0.986	0.947	73
PortScan	0.814	0.997	0.896	1607
SSH-Patator	0.649	1	0.788	63
Web Attack - Brute Force	0.022	0.071	0.033	14
Web Attack - XSS	0.054	0.833	0.101	6
Accuracy	**N/A**	**N/A**	**0.959**	28279
Macro Average	0.639	0.9	0.7	28279
Weighted Average	0.977	0.96	0.97	28279

As evidenced by the metrics, high-volume volumetric intrusions achieve exceptional recall values exceeding **0.995**, validating the efficacy of the sequence-learning mechanisms. Conversely, severe minority categories like *Web Attack - Brute Force* exhibit lower precision due to overlapping syntactic feature spaces with legitimate HTTP transactions.

DOI: *10.5281/zenodo.22333206*

5.2 Confusion Matrix Analysis

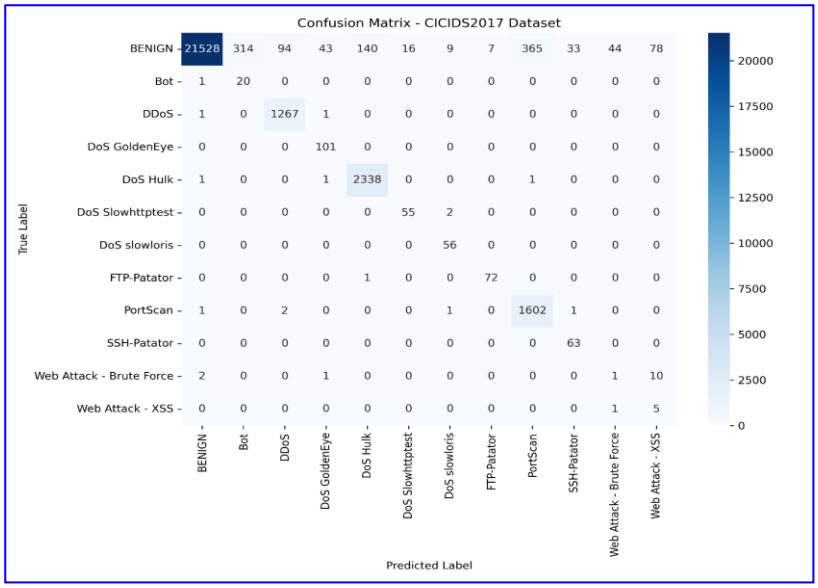

Fig. 1. Confusion matrix of the hybrid CNN-LSTM model on CICIDS2017

The confusion matrix provides granular insight into class-specific performance:

Volumetric attacks such as DDoS (1,267 correct), DoS Hulk (2,338 correct), and PortScan (1,602 correct) were classified with exceptional precision [4].

Minor misclassifications occurred where 314 BENIGN flows were incorrectly predicted as Bot traffic, highlighting the overlap between legitimate background applications and automated heartbeats [4, 8].

5.3 ROC Curve and AUC Analysis

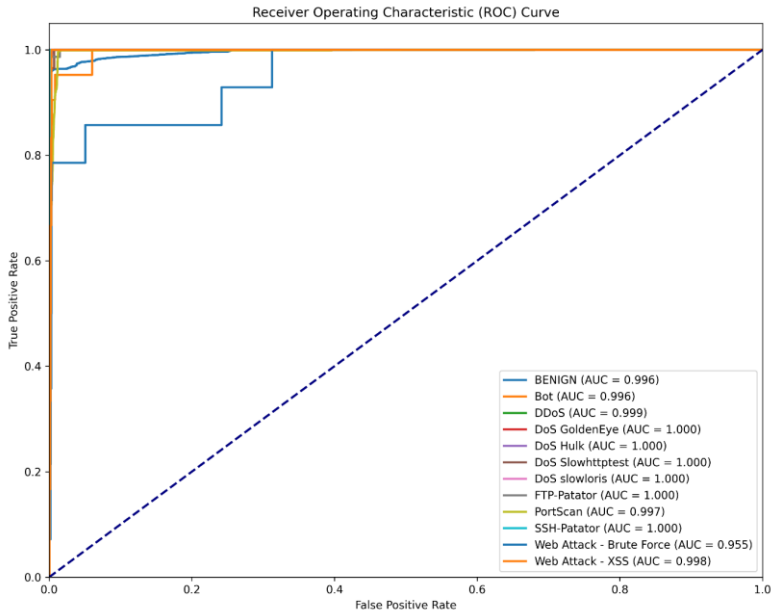

Fig. 2. ROC curves of the hybrid CNN-LSTM model on CICIDS2017

 DOI: *10.5281/zenodo.22333206*

The ROC-AUC curves confirm the superior discriminative capability of the model:

Flawless AUC scores of 1.000 were achieved for DoS GoldenEye, DoS Hulk, DoS Slowhttptest, and FTP-Patator [4].

Challenging classes like Web Attack - Brute Force attained a strong AUC of 0.955, proving that the model possesses high underlying confidence even when hard-label thresholds require fine-tuning [3, 4].

5.4 Model Explainability via SHAP

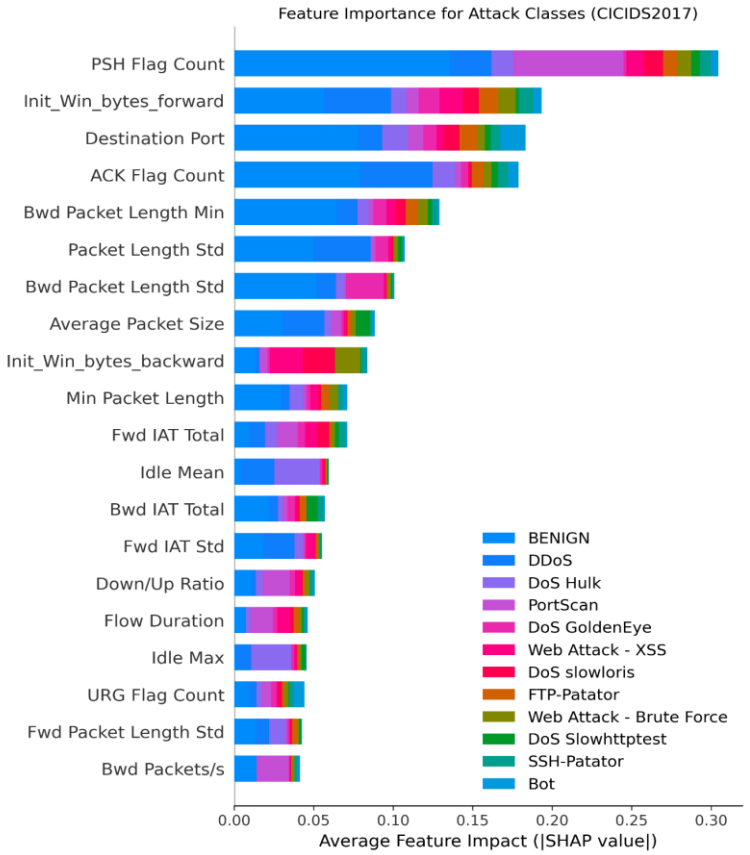

Fig. 3. SHAP feature importance summary for attack classes on CICIDS2017

The SHAP summary bar plot elucidates the global feature importance rankings:

PSH Flag Count, Destination Port, ACK Flag Count, and Init_Win_bytes_forward emerged as the top-ranking discriminative features [4, 9].

This aligns perfectly with networking principles, as DoS/DDoS attacks rely heavily on manipulating TCP push and acknowledgment flags [4, 9]. This explainability directly reinforces human trust management within SOC environments [12, 13].

 DOI: *10.5281/zenodo.22333206*

5.5 Multi-Context Feature Shift and Concept Drift

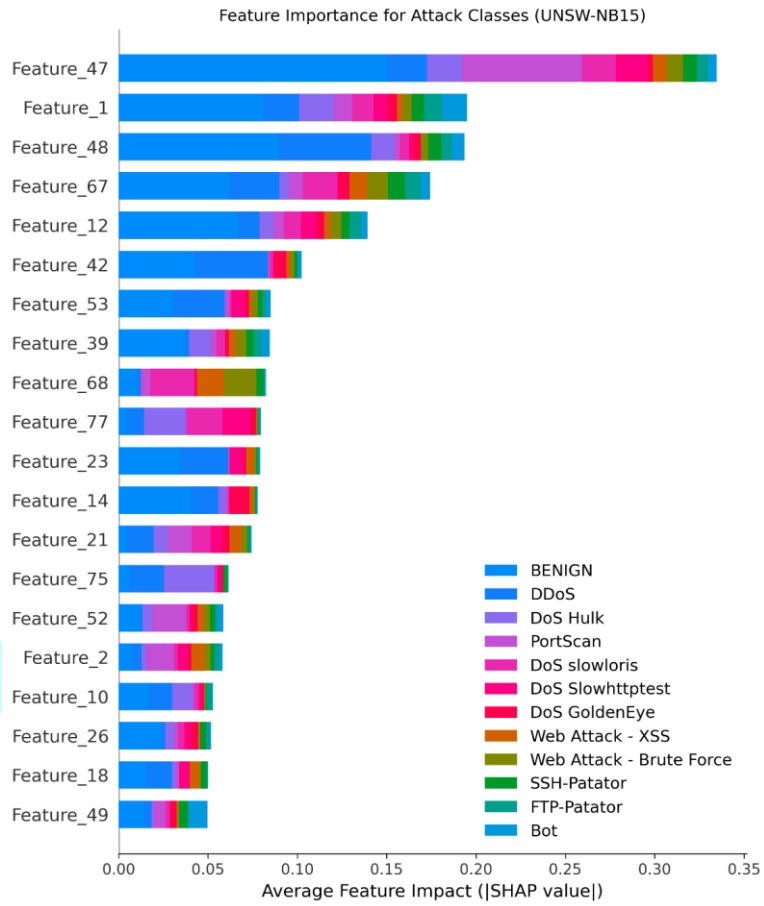

Fig. 4. SHAP feature importance summary for attack classes on the UNSW-NB15 dataset

To empirically validate the model's resilience against concept drift, a secondary SHAP analysis was conducted on the UNSW-NB15 dataset (**Fig. 4**) [2]. A comparative evaluation between **Fig. 3** and **Fig. 4** visually demonstrates a pronounced multi-context feature shift [4, 5]. While the model prioritizes protocol flags (e.g., PSH Flag Count, ACK Flag Count) within the CICIDS2017 topology (**Fig. 3**), its attention dynamically reweights toward structural flow indices and indices such as Feature_47, Feature_1, and Feature_67 when evaluated on UNSW-NB15 (**Fig. 4**). This adaptive feature redistribution confirms that the hybrid CNN-LSTM architecture successfully captures underlying contextual behaviors rather than over-fitting to dataset-specific artifacts.

6. Conclusion

By combining deep learning performance with rigorous explainability, adversarial auditing, **and cross-dataset XAI visual validation**, the proposed system successfully

DOI: *10.5281/zenodo.22333206*

overcomes the black-box dilemma, significantly enhancing operational trust management for next-generation cybersecurity infrastructures.

Acknowledgments. We acknowledge Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam for supporting this study.

Disclosure of Interests. The author has no conflicts of interest regarding the content of this article.

7. References

1. A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018.
2. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A Survey on Concept Drift Adaptation," ACM Computing Surveys, vol. 46, no. 4, Article 44, 2014.
3. S. Naderi Mighan and M. Kahani, "A novel scalable intrusion detection system based on deep learning," International Journal of Information Security, vol. 20, pp. 387–403, 2021.
4. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "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), pp. 108–116, 2018.
5. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems (UNSW-NB15 Network Data Set)," in Proceedings of the Military Communications and Information Systems Conference (MilCIS), pp. 1–6, 2015.
6. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

 DOI: *10.5281/zenodo.22333206*

7. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.

8. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp. 41525–41550, 2019.

9. M. Wang, K. Zheng, Y. Yang, and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," IEEE Access, vol. 8, pp. 73127–73141, 2020.

10. M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," Journal of Information Security and Applications, vol. 50, p. 102246, 2020.

11. D. Dasgupta, Z. Akhtar, and S. Sen, "Machine learning in cybersecurity: A comprehensive survey," Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, vol. 19, no. 1, pp. 1–50, 2020.

12. D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An Adversarial Approach for Explainable AI in Intrusion Detection Systems," in Proceedings of IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, pp. 3237–3243, 2018.

13. B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model," Complexity, vol. 2021, Article ID 6634811, 2021.

14. "Network Intrusion Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset.

15. "UNSW-NB15 Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15.

 DOI: *10.5281/zenodo.22333206*