Journal of
Cybersecurity and Privacy

MDPI

Editorial

Diverse Intrusion and Malware Detection: AI-Based and Non-AI-Based Solutions

Feng Wang [1,*,†] and Yongning Tang [2,†]

[1] School of Engineering, Liberty University, Lynchburg, VA 24515, USA
[2] School of Information Technology, Illinois State University, Normal, IL 61761, USA; ytang@ilstu.edu
* Correspondence: fwang@liberty.edu
† These authors contributed equally to this work.

In today's interconnected world, the need for robust intrusion and malware detection and prevention has never been more critical. Despite significant advances in developing intrusion detection systems (IDS), achieving holistic security continues to be an evolving challenge. The sophistication of cyber threats necessitates IDS models that are both robust and adaptable.

In 2010, a seminal work by Robin Sommer and Vern Paxson discussed the challenges of applying machine learning (ML) for anomaly detection [1]. They highlighted limitations such as imbalanced training cases and high false-positive rates, as well as the importance of feature selection. In response, researchers have explored various ML algorithms to enhance IDS for detecting cyber attacks, such as spam classification [2,3], malware detection [4], and intrusion detection [5–8].

Recent years have witnessed the rise of Deep Learning (DL) methods as powerful tools in detecting cyber threats [9–14]. Various ML/DL models have been leveraged, including autoencoder [11,15], LSTM [16], recurrent neural networks (RNNs) [14,17], Bayesian [12], Feedforward Deep Neural Network (FFDNN) [13], Convolutional Neural Network (CNN) and LSTM [18], and GRU [19]. These studies consistently demonstrate that ML and DL models significantly enhance the accuracy of detecting normal and anomalous traffic [9].

Since AI models require vast amounts of data for training, to effectively utilize these models in intrusion detection systems, we need to address two main challenges: feature selection and data imbalance [1].

Feature selection is the process of identifying the most relevant features for use in building a machine learning model. The accuracy of ML-based methods is heavily influenced by the quality of the feature space [20–23]. Therefore, developing efficient feature selection techniques is crucial for optimizing detection accuracy.

Data imbalance is another significant challenge. IDS datasets often contain far fewer examples of malicious traffic compared to normal traffic. Training models on such imbalanced datasets can lead to poor detection performance [24]. Techniques such as random oversampling and the Synthetic Minority Oversampling Technique (SMOTE) are commonly used to create balanced datasets from imbalanced data, thereby improving model accuracy [25–27].

While AI-based detection models have made significant strides, there remains a crucial need for robust non-AI-based solutions. The opaque nature of AI models, often referred to as "black boxes", can make it difficult to understand their decision-making processes. Non-AI approaches, based on well-defined rules and logic, provide greater explainability and transparency. These solutions can complement AI-based defenses by offering clear and understandable detection mechanisms, which are particularly valuable in scenarios like malware detection and identifying malicious behaviors.

A comprehensive IDS should leverage a diverse set of tools, incorporating both AI and non-AI solutions. By understanding and integrating the strengths and limitations of each

check for updates

Citation: Wang, F.; Tang, Y. Diverse Intrusion and Malware Detection: AI-Based and Non-AI-Based Solutions. *J. Cybersecur. Priv.* **2024**, *4*, 382–387. https://doi.org/10.3390/jcp4020019

Received: 24 May 2024
Revised: 1 June 2024
Accepted: 6 June 2024
Published: 11 June 2024

approach, we can develop a more resilient and effective defense against the continuously evolving threat landscape.

This Special Issue [28] is dedicated to advancing the field of intrusion and malware detection across various network environments, including future Internet architectures, 5G and beyond wireless networks, enterprises, data centers, edge and cloud networks, software-defined networking (SDN), optical networks, and IoT-scale networks. From the 23 manuscripts submitted to this Special Issue, 10 were rigorously reviewed and accepted for publication. These contributions, listed below, reflect the diverse and innovative approaches in both AI and non-AI realms.

1. Wang, F.; Tang, Y.; Fang, H. Mitigating IoT Privacy-Revealing Features by Time Series Data Transformation. *J. Cybersecur. Priv.* **2023**, *3*, 209–226. https://doi.org/10.3390/jcp3020012.
2. Li, R.; Tsikerdekis, M. Hourly Network Anomaly Detection on HTTP Using Exponential Random Graph Models and Autoregressive Moving Average. *J. Cybersecur. Priv.* **2023**, *3*, 435–450. https://doi.org/10.3390/jcp3030022.
3. Ghani, H.; Virdee, B.; Salekzamankhani, S. A Deep Learning Approach for Network Intrusion Detection Using a Small Features Vector. *J. Cybersecur. Priv.* **2023**, *3*, 451–463. https://doi.org/10.3390/jcp3030023.
4. Ahmadi Abkenari, F.; Milani Fard, A.; Khanchi, S. Hybrid Machine Learning-Based Approaches for Feature and Overfitting Reduction to Model Intrusion Patterns. *J. Cybersecur. Priv.* **2023**, *3*, 544–557. https://doi.org/10.3390/jcp3030026.
5. Abdelmoumin, G.; Rawat, D.; Rahman, A. Studying Imbalanced Learning for Anomaly-Based Intelligent IDS for Mission-Critical Internet of Things. *J. Cybersecur. Priv.* **2023**, *3*, 706–743. https://doi.org/10.3390/jcp3040032.
6. Ghani, H.; Salekzamankhani, S.; Virdee, B. A Hybrid Dimensionality Reduction for Network Intrusion Detection. *J. Cybersecur. Priv.* **2023**, *3*, 830–843. https://doi.org/10.3390/jcp3040037.
7. Ghosh, T.; Bagui, S.; Bagui, S.; Kadzis, M.; Bare, J. Anomaly Detection for Modbus over TCP in Control Systems Using Entropy and Classification-Based Analysis. *J. Cybersecur. Priv.* **2023**, *3*, 895–913. https://doi.org/10.3390/jcp3040041.
8. Rose, A.; Graham, S.; Schubert Kabban, C.; Krasnov, J.; Henry, W. ScriptBlock Smuggling: Uncovering Stealthy Evasion Techniques in PowerShell and .NET Environments. *J. Cybersecur. Priv.* **2024**, *4*, 153–166. https://doi.org/10.3390/jcp4020008.
9. Halder, R.; Das Roy, D.; Shin, D. A Blockchain-Based Decentralized Public Key Infrastructure Using the Web of Trust. *J. Cybersecur. Priv.* **2024**, *4*, 196–222. https://doi.org/10.3390/jcp4020010.
10. Muhati, E.; Rawat, D. Data-Driven Network Anomaly Detection with Cyber Attack and Defense Visualization. *J. Cybersecur. Priv.* **2024**, *4*, 241–263. https://doi.org/10.3390/jcp4020012.

As shown in Table 1, this Special Issue addresses a wide spectrum of topics within intrusion and malware detection and prevention. These topics include feature reduction, feature selection, handling imbalanced data, addressing new threats, ensuring data privacy, and introducing new infrastructure for public key infrastructure (PKI). The proposed methods are evaluated using a variety of datasets, ensuring robust and comprehensive analysis. The majority of the contributions (1, 2, 3, 4, 5, 6, 7) focus on AI-based intrusion and malware detection and data privacy, while three specifically explore non-AI-based solutions. This balanced approach underscores the importance of integrating both AI and non-AI methodologies to develop more effective and transparent intrusion detection systems.

IDSs rely on large feature sets, but some features contain irrelevant and redundant information, which increases computational complexity and decreases accuracy. There are four papers in this Special Issue that address feature selection, feature reduction, and overfitting.

Table 1. Analysis of the published contributions in the Special Issue.

Contribution #	Research Problem	Detection Models	Datasets
1	Protect the privacy of IoT devices	LSTM	BoT-IoT and UNSW-NB15
2	Detect and prevent data exfiltration	Exponential random graph models	University of New Brunswick's ISCX 2012
3	Feature selection and feature extraction	Ensemble of Support Vector (SVC), K-Nearest Neighbor (KNN), and Deep Neural Network (DNN)	UNSW-NB15
4	Data imbalance	PCA and oSVM	BoT-IoT
5	Feature reduction and overfitting	Decision Tree, Linear Regression, Boruta, Random Forest, LASSO, and autoencoders	CSE-CIC-IDS2018
6	Feature reduction and feature selection	Feedforward Neural Network (FFNN)	UNSW-NB15 and NSL-KDD
7	Feature selection	BayesNet, Naïve Bayes, J48, Simple Logistic, SVM, Multilayer Perceptron, Random Forest, and Decision Table	Modbus over TCP/IP Data
8	ScriptBlock Smuggling	Malware Detection	N/A
9	Decentralized public key infrastructure	Web of Trust (WoT) and blockchain	N/A
10	Network compromises and malware patterns	Visualization	KDDCUP'99

Ghani et al., in their first paper, propose a hybrid dimensionality reduction system that combines feature selection and feature extraction. They employ the Recursive Feature Elimination (RFE) technique to identify and eliminate irrelevant or redundant features from the initial dataset. Subsequently, they use Principal Component Analysis (PCA) to transform the remaining features into a lower-dimensional representation while preserving the most important information. Their system successfully reduces the original 41 features to a more manageable set of 15 components. Importantly, the classification performance, using an ensemble of Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), and Deep Neural Network (DNN) classifiers, remains robust, indicating that the reduced and transformed features do not significantly compromise the system's ability to detect network intrusions compared to using the full feature set.

In their second paper, Ghani et al. propose a deep learning-based approach for network intrusion detection utilizing a Feedforward Neural Network (FFNN). They focus on achieving high classification accuracy with a reduced feature vector. Their approach demonstrates that a smaller, more targeted feature vector can be equally effective in detecting network traffic anomalies within datasets like UNSW-NB15 and NSL-KDD. This not only improves classification accuracy but also reduces the computational power required for analysis.

Ghosh et al. propose a statistical approach utilizing cluster-based entropy analysis on selected network traffic features. They focus on features such as packet size, inter-packet interval, packet process time, and two additional Modbus application protocol header features: Modbus frame length and function code value. Their classification-based analysis reveals that incorporating the two Modbus-specific features along with the three TCP/IP features significantly improves classification accuracy for DOS attacks compared to MITM attacks.

Ahmadi et al. address the challenges of feature reduction and model overfitting in network intrusion detection. They conduct experiments using a subset of the CSE-CIC-2018 dataset, evaluating various feature reduction approaches, including Linear Regression, Boruta, Random Forest with IncMSE, Random Forest with IncNodePurity, LASSO, and autoencoders. To assess the effectiveness of each approach in mitigating overfitting, they calculate the Root-Mean-Squared Error (RMSE) between the training and testing datasets for each model combined with a Decision Tree classifier. Their findings reveal that the combination of a Decision Tree classifier and features reduced using autoencoders achieves the lowest RMSE, indicating the most effective reduction in overfitting among all the tested scenarios.

Data imbalance is another significant challenge in training machine learning models for IDS. Imbalanced datasets occur when there are significantly fewer examples of malicious attacks compared to normal network traffic, leading to detection inaccuracies. One paper specifically focuses on addressing this issue.

Abdelmoumin et al. investigate the impact of various techniques on data imbalance. Their work focuses on three main approaches: oversampling (increasing minority class examples), undersampling (decreasing majority class examples), and generating new synthetic samples for the minority class using generative methods. They evaluate these techniques by analyzing their impact on the performance and prediction accuracy of the models. They measure how well the trained models perform on balanced datasets compared to imbalanced ones, and assess the robustness of the models to new attacks that share similarities with existing ones. By investigating these techniques, Abdelmoumin et al. aim to identify the most effective methods for mitigating data imbalance and improving the overall performance and robustness of machine learning-based IDS.

There are two papers introducing noticeable threats: ScriptBlock Smuggling and data exfiltration.

ScriptBlock Smuggling is a novel threat that manipulates PowerShell and .NET environments to bypass the Antimalware Scan Interface (AMSI) on Windows operating systems. AMSI is crucial for malware detection, but ScriptBlock Smuggling exploits vulnerabilities to evade it. This threat hinges on manipulating ScriptBlocks, which are fundamental units of PowerShell code. By altering ScriptBlocks within their Abstract Syntax Tree (AST), attackers can create a dual representation. One representation caters to the compiler for normal execution, while the other is specifically designed to deceive antivirus software and log analysis tools. This allows malicious code to bypass AMSI detection and renders traditional memory patching bypass methods ineffective.

The research by Rose et al. delves into the inner workings of ScriptBlock creation within PowerShell, analyzes its built-in security features, and exposes critical limitations in AMSI's ability to scrutinize ScriptBlocks effectively. Furthermore, it explores the implications of log spoofing as an integral part of this evasion method. These findings highlight potential avenues for attackers to exploit these weaknesses, suggesting the emergence of a new class of techniques to bypass AMSI and manipulate logs. To address this growing threat, the paper proposes a synchronization strategy for ASTs, aiming to unify the processes of code compilation and malware scanning. This strategy could ultimately reduce the attack surface within PowerShell and .NET environments.

Data exfiltration is a cyberattack where unauthorized individuals steal or copy sensitive information. Examples include credit card numbers, Social Security numbers, and personal details exposed in the 2017 Equifax breach, where the data of 143 million Americans was compromised [29]. Li et al. propose a novel approach to detect data exfiltration using network graphs. Their method leverages the concept of network topology, which maps the connections and data flow within a network. Then, the topology information is incorporated in a statistical model to detect anomalies. More specifically, hourly HTTP data are aggregated to construct graphs. Nodes represent source and destination IP addresses, while edges represent the total byte volume transferred between them. Nodes are then categorized as servers or hosts based on their port numbers, resulting in bipartite graphs. Exponential random graph models (ERGMs) are employed to convert the network's topological features into a time series. Finally, Autoregressive Moving Average (ARMA) is used to identify deviations in the time series, potentially indicating malicious exfiltration attempts. This approach offers valuable insights into network behavior and can aid cybersecurity analysts in making informed decisions alongside existing intrusion detection systems.

Muharti and Rawat propose a data analytics-driven network anomaly detection model uniquely complemented by a visualization layer, providing real-time insights into cyber attacks and their defenses. This approach utilizes network scanning tools and discovery services to visualize the network by identifying live IP-based devices. A data analytics-based intrusion detection system scrutinizes all network connections, and mitigation measures are initiated by visually distinguishing malicious from benign connections using red and blue hues, respectively.

One paper addresses and mitigates the vulnerabilities of centralized certificate verification using blockchain technology. The centralization of existing PKI systems introduces significant vulnerabilities, as a compromised CA can issue unauthorized certificates and access sensitive information. Halder et al. address and mitigate these vulnerabilities through decentralized certificate verification using blockchain technology. They present a decentralized public key infrastructure (PKI) based on a distributed trust model, such as the Web of Trust (WoT) and blockchain technologies, to overcome issues like single points of failure and prevent tampering with existing certificates. Additionally, their infrastructure establishes a trusted key-ring network that decouples the authentication process from CAs, enhancing secure certificate issuance and accelerating the revocation process. Their experimental results demonstrate the effectiveness of this proposed system in practice, despite incurring additional overhead compared to conventional PKIs.

Finally, there is one paper addressing IoT data privacy leakage. Wang et al. consider the challenge of protecting the privacy of IoT devices by transforming time series datasets. The transformed datasets retain the intrinsic value of the original IoT data while maintaining data utility. This approach enables non-expert data owners to better understand and evaluate the potential device-level privacy risks associated with their IoT data, while simultaneously offering a reliable solution to mitigate their concerns about privacy violations.

Author Contributions: Conceptualization, F.W. and Y.T.; methodology, F.W.; validation, F.W. and Y.T.; formal analysis, F.W.; investigation, F.W. and Y.T.; data curation, F.W.; writing—original draft preparation, F.W.; writing—review and editing, Y.T.; visualization, F.W. and Y.T.; supervision, Y.T.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Sommer, R.; Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy, Berkeley/Oakland, CA, USA, 16–19 May 2010; pp. 305–316.
2. Dada, E.G.; Bassi, J.S.; Chiroma, H.; Adetunmbi, A.O.; Ajibuwa, O.E. Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems. *Heliyon* **2019**, *5*, e01802. [CrossRef] [PubMed]
3. Jain, A.K.; Goel, D.; Agarwal, S.; Singh, Y.; Bajaj, G. Predicting Spam Messages using Back Propagation Neural Network. *Wirel. Pers. Commun.* **2020**, *110*, 403–422. [CrossRef]

4. Ma, Z.; Ge, H.; Liu, Y.; Zhao, M.; Ma, J. A Combination Method for Android Malware Detection based on Control Flow Graphs and Machine Learning Algorithms. *IEEE Access* **2019**, *7*, 21235–21245. [CrossRef]
5. Gharaee, H.; Hosseinvand, H. A New Feature Selection IDS based on Genetic Algorithm and SVM. In Proceedings of the 2016 8th International Symposium on Telecommunications (IST), Tehran, Iran, 27 August 2016; pp. 139–144.
6. Belouch, M.; Hadaj, S.E.; Idhammad, M. A Two-Stage Classifier Approach using Reptree Algorithm for Network Intrusion Detection. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*. [CrossRef]
7. Dey, S.; Ye, Q.; Sampalli, S. A Machine Learning based Intrusion Detection Scheme for Data Fusion in Mobile Clouds Involving Heterogeneous Client Networks. *Inf. Fusion* **2019**, *49*, 205–215. [CrossRef]
8. Mirsky, Y.; Doitshman, T.; Elovici, Y.; Shabtai, A. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. *arXiv* **2018**, arXiv:1802.09089.
9. Kilincer, I.F.; Fatih, E.; Abdulkadir, S. Machine Learning Methods for Cyber Security Intrusion Detection: Datasets and Comparative Study. *Comput. Netw.* **2021**, *188*, 107840. [CrossRef]
10. Xin, Y.; Kong, L.; Liu, Z.; Chen, Y.; Li, Y.; Zhu, H.; Wang, C. Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access* **2018**, *6*, 35365–35381. [CrossRef]
11. Hijazi, A.; El Safadi, A.; Flaus, J.M. A Deep Learning Approach for Intrusion Detection System in Industry Network. In Proceedings of the BDCSIntell, Hadath, Lebanon, 13–15 December 2018; pp. 55–62.
12. Sivanathan, A.; Gharakheili, H.H.; Loi, F.; Radford, A.; Wijenayake, C.; Vishwanath, A.; Sivaraman, V. Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics. *IEEE Trans. Mob. Comput.* **2019**, *18*, 1745–1759. [CrossRef]
13. Mambwe, K.S.; Sun, Y. A Deep Learning Method with Wrapper based Feature Extraction for Wireless Intrusion Detection System. *Comput. Secur.* **2020**, *92*, 101752.
14. Mambwe, K.S. A Deep Learning Technique for Intrusion Detection System using a Recurrent Neural Networks based Framework. *Comput. Commun.* **2023**, *199*, 113–125.
15. Ali H.M.; Cosan, S. Computer Network Intrusion Detection using Sequential LSTM Neural Networks Autoencoders. In Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.
16. Hong, D.R.; Li, X.Y.; Zhang, Q.Y.; Yuan, H. Network Intrusion Detection Model based on Multivariate Correlation Analysis—Long Short—Time Memory Network. *IET Inf. Secur.* **2020**, *14*, 166–174.
17. Yin, C.; Zhu, Y.; Fei, J.; He, X. A Deep Learning Approach for Intrusion Detection using Recurrent Neural Networks. *IEEE Access* **2017**, *5*, 21954–21961. [CrossRef]
18. Hsu, C.M.; Hsieh, H.Y.; Prakosa, S.W.; Azhari, M.Z.; Leu, J.S. Using long-short-term memory based convolutional neural networks for network intrusion detection. In Proceedings of the Wireless Internet: 11th EAI International Conference, WiCON 2018, Taipei, Taiwan, 15–16 October 2018; pp. 86–94.
19. Li, Z.; Rios AL, G.; Xu, G.; Trajković, L. Machine Learning Techniques for Classifying Network Anomalies and Intrusions. In Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS 2019, Sapporo, Japan, 26–29 May 2019; pp. 1–5.
20. Salo, F.; Nassif, A.B.; Essex, A. Dimensionality Reduction with IGPCA and Ensemble Classifier for Network Intrusion Detection. *Comput. Netw.* **2019**, *148*, 164–175. [CrossRef]
21. Zebari, R.; Abdulazeez, A.; Zeebaree, D.; Zebari, D.; Saeed, J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 56–70. [CrossRef]
22. Chaouki, K.; Krichen, S. A NSGA2-LR Wrapper Approach for Feature Selection in Network Intrusion Detection. *Comput. Netw.* **2020**, *172*, 107183.
23. Injadat, M.; Moubayed, A.; Nassif, A.B.; Shami, A. Multi-stage Optimized Machine Learning Framework for Network Intrusion Detection. *IEEE Trans. Netw. Serv. Manag.* **2020**, *18*, 1803–1816. [CrossRef]
24. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]
25. Ahmed, S.; Mahbub, A.; Rayhan, F.; Jani, R.; Shatabda, S.; Farid, D.M. Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques. In Proceedings of the 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 21–23 December 2017; pp. 1–5.
26. Johnson, J.M.; Khoshgoftaar, T.M. Survey on Deep Learning with Class Imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
27. Desuky, A.S.; Hussain, S. An Improved Hybrid Approach for Handling Class Imbalance Problem. *Arab. J. Sci. Eng.* **2021**, *46*, 3853–3864. [CrossRef]
28. Intrusion, Malware Detection and Prevention in Networks. Available online: https://www.mdpi.com/journal/jcp/special_issues/U21OHBD667 (accessed on 31 May 2024).
29. Jeff, L. Apache Struts 2: How Technical and Development Gaps Caused the Equifax Breach. *Netw. Secur*. **2018**, *1*, 5–8.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.