10.48047/jocaaa.2024.33.08.455

An Explainable AI-Driven Intrusion Detection System Using Optimized Gradient Boosting for Real-Time Cyber Threat Intelligence

Ritu Raj Kashyap, Abhay Shukla, Somendra Tripathi
Department of Computer Science and Engineering, FET, Rama University Uttar Pradesh Kanpur 209217
Email id:kashyaprituraj76@gmail.com

Abstract

The rapid evolution of cyber threats has significantly increased the demand for intelligent, adaptive, and interpretable intrusion detection systems (IDS). Traditional intrusion detection approaches, particularly signature-based systems, are limited in detecting zero-day and evolving cyber-attacks. Although machine learning-based IDS frameworks have demonstrated improved detection capabilities, their lack of interpretability restricts their deployment in critical cybersecurity infrastructures where transparency and accountability are essential.

This paper proposes an Explainable Artificial Intelligence (XAI)-driven intrusion detection framework that integrates an optimized Extreme Gradient Boosting (XGBoost) classifier with SHapley Additive exPlanations (SHAP) to achieve both high predictive performance and interpretability. The proposed model is evaluated using a cleaned and preprocessed version of the CICIDS2017 dataset, which provides realistic network traffic scenarios with reduced noise and redundancy.

Experimental results demonstrate that the model achieves an accuracy of 99.88%, precision of 99.37%, recall of 99.96%, F1-score of 0.996, and a ROC-AUC score of 0.9999. These results confirm the model's ability to accurately distinguish between benign and malicious traffic while maintaining a high degree of reliability. Furthermore, SHAP-based explainability provides meaningful insights into feature contributions, enabling cybersecurity analysts to understand model decisions.

The proposed framework effectively bridges the gap between high-performance intrusion detection and model interpretability, making it suitable for real-time cyber threat intelligence applications.

Keywords: Intrusion Detection System, Explainable Artificial Intelligence, XGBoost, SHAP, Cybersecurity, Machine Learning, CICIDS2017, Network Security

1.Introduction

The exponential growth of digital communication systems and internet-based services has led to a corresponding increase in cybersecurity threats. Modern networks are continuously exposed to various forms of attacks, including Distributed Denial of Service (DDoS), brute-force attacks, infiltration, and data exfiltration. These threats pose significant risks to organizations, making effective network monitoring and protection essential. Intrusion Detection Systems (IDS) are designed to monitor network traffic and detect malicious activities. Traditionally, IDS techniques are classified into signature-based and anomaly-based approaches. Signature-based IDS rely on predefined attack patterns and are effective in detecting known threats; however, they fail to identify novel or zero-day attacks. Anomaly-based IDS, on the other hand, detect deviations from normal behavior and can identify new threats, but they often generate a high number of false positives. To overcome these limitations, machine learning (ML) techniques have been widely adopted in intrusion detection. ML-based IDS can learn complex patterns from data and generalize to unseen attack scenarios. However, most ML models operate as black-box systems, providing limited insight into their internal decision-making processes. This lack of interpretability is a critical drawback, particularly in cybersecurity applications where understanding the rationale behind predictions is essential for trust and decision-making. Explainable Artificial Intelligence (XAI) has merged as a promising solution to address this challenge. XAI techniques aim to provide transparent

10.48047/jocaaa.2024.33.08.455

and interpretable explanations for model predictions. Among these techniques, SHAP has gained significant attention due to its strong theoretical foundation and consistency in feature attribution. In this study, we propose an explainable intrusion detection system that combines the predictive power of XGBoost with SHAP-based interpretability. The proposed model is trained and evaluated on the CICIDS2017 dataset, which provides realistic network traffic data. The primary objective is to develop a system that achieves high detection accuracy while offering interpretable insights, thereby enhancing its applicability in real-world cybersecurity environments.

2.Related Work

Intrusion detection has been extensively studied using a wide range of machine learning and deep learning approaches. Early research focused on traditional algorithms such as Decision Trees, Naïve Bayes, and Support Vector Machines, which provided moderate performance in detecting known attack patterns. Ensemble learning methods, such as Random Forest, improved detection accuracy by combining multiple classifiers and reducing variance.

With advancements in computational capabilities, deep learning techniques have been introduced for intrusion detection. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated the ability to capture complex patterns in network traffic data. However, these models are computationally expensive and often lack interpretability, limiting their practical deployment.

Recent research has shifted toward integrating explainability into machine learning models. Techniques such as LIME and SHAP have been widely used to provide post-hoc explanations for model predictions. SHAP, in particular, has gained prominence due to its solid theoretical basis in cooperative game theory and its ability to provide consistent feature importance values.

XGBoost has emerged as a highly effective algorithm for structured data due to its scalability, efficiency, and regularization capabilities. Several studies have demonstrated the effectiveness of XGBoost in intrusion detection tasks. However, limited research has focused on combining XGBoost with explainable AI techniques to enhance both performance and interpretability.

This paper contributes to the existing literature by proposing an explainable IDS framework that integrates XGBoost with SHAP, achieving high detection accuracy while maintaining transparency.

3.Methodology

3.1 Dataset Description

The CICIDS2017 dataset is used in this study as it provides a comprehensive and realistic representation of network traffic. The dataset includes both benign traffic and various types of cyber-attacks, making it suitable for evaluating intrusion detection systems.

A cleaned and preprocessed version of the dataset is utilized to ensure data quality. The preprocessing process removes duplicate records, handles missing and infinite values, and eliminates redundant features through correlation analysis. The final dataset consists of 53 numerical features representing network flow characteristics.

The target variable is converted into a binary classification problem, where normal traffic is labeled as 0 and malicious traffic as 1.

3.2 Data Preprocessing

Data preprocessing plays a crucial role in ensuring the reliability and performance of machine learning models. In this study, infinite values are replaced with null values and subsequently removed. The dataset is then split into training and testing sets using stratified sampling to preserve class distribution.

Feature scaling is applied using standardization techniques to normalize the feature space. Importantly, scaling is performed after the train-test split to prevent data leakage and ensure unbiased model evaluation.

3.3 Model Development

10.48047/jocaaa.2024.33.08.455

The proposed model is based on the XGBoost algorithm, which utilizes gradient boosting to build an ensemble of decision trees. Each tree is trained to correct the errors of the previous ones, resulting in a highly accurate model.

The objective function of XGBoost is defined as:

[
\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}*i) + \sum*{k=1}^{K} \Omega(f_k)
]

where (l) represents the loss function and (\Omega) is the regularization term that controls model complexity.

To address class imbalance, the scale_pos_weight parameter is introduced, which adjusts the importance of positive and negative samples during training. Hyperparameters such as learning rate, number of estimators, and tree depth are carefully tuned to optimize performance.

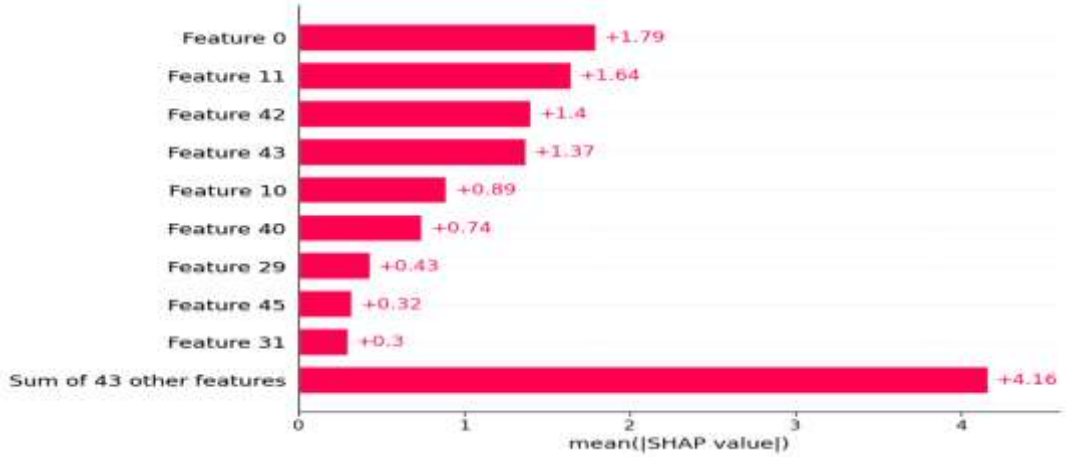
3.4 Explainability Framework

SHAP is employed to interpret the predictions of the XGBoost model. It assigns each feature a contribution value based on its impact on the prediction outcome. SHAP provides both global and local explanations, enabling a comprehensive understanding of model behavior.

The SHAP value is calculated as:

[
\phi_i = \sum_{S \subseteq F \setminus {i}}
\frac{|S|!(|F|-|S|-1)!}{|F|!}
\left[f(S \cup {i}) - f(S) \right]
]

Fig 1: SHAP-based global feature importance of the proposed XGBoost model. The plot highlights the most influential network traffic features contributing to intrusion detection, demonstrating the model's ability to identify key indicators of malicious activity.



4.Experimental Results

4.1 Performance Metrics

The performance of the proposed model is evaluated using standard metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

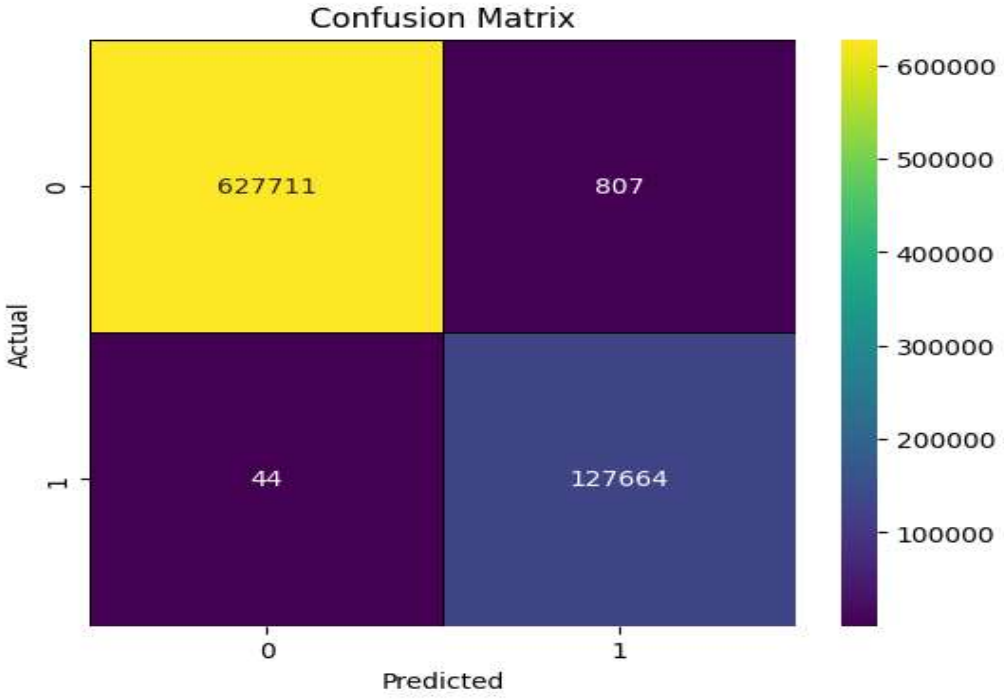
4.2 Results

T The experimental results demonstrate the effectiveness of the proposed model:

Metric	Value
Accuracy	99.88%
Precision	99.37%
Recall	99.96%
F1 Score	0.996

10.48047/jocaaa.2024.33.08.455

ROC-AUC 0.9999



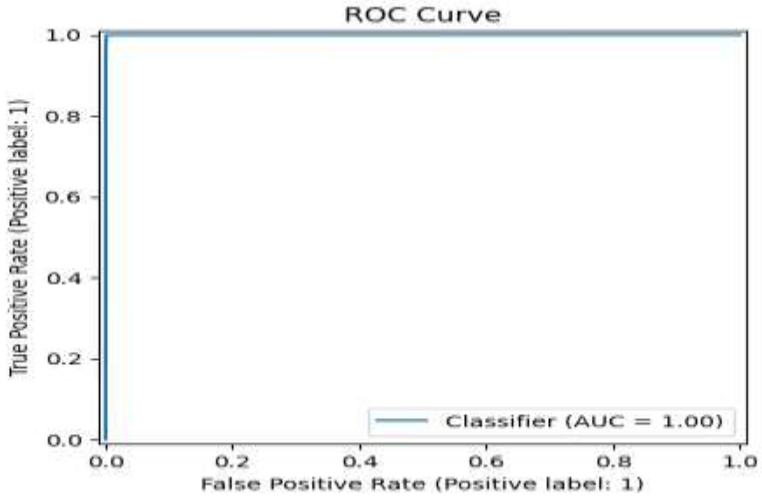
4.3 Confusion Matrix

Fig 2: Confusion matrix of the proposed intrusion detection system. The results indicate high true positive and true negative rates with minimal false classifications, confirming the robustness and reliability of the model.

The confusion matrix indicates that the model achieves high true positive and true negative rates with minimal misclassification. This demonstrates the robustness and reliability of the proposed intrusion detection system.

4.4 ROC Curve

Fig 3: Receiver Operating Characteristic (ROC) curve of the proposed model. The near-perfect area under the curve (AUC) demonstrates the model's strong discriminative capability in distinguishing between benign and malicious network traffic.

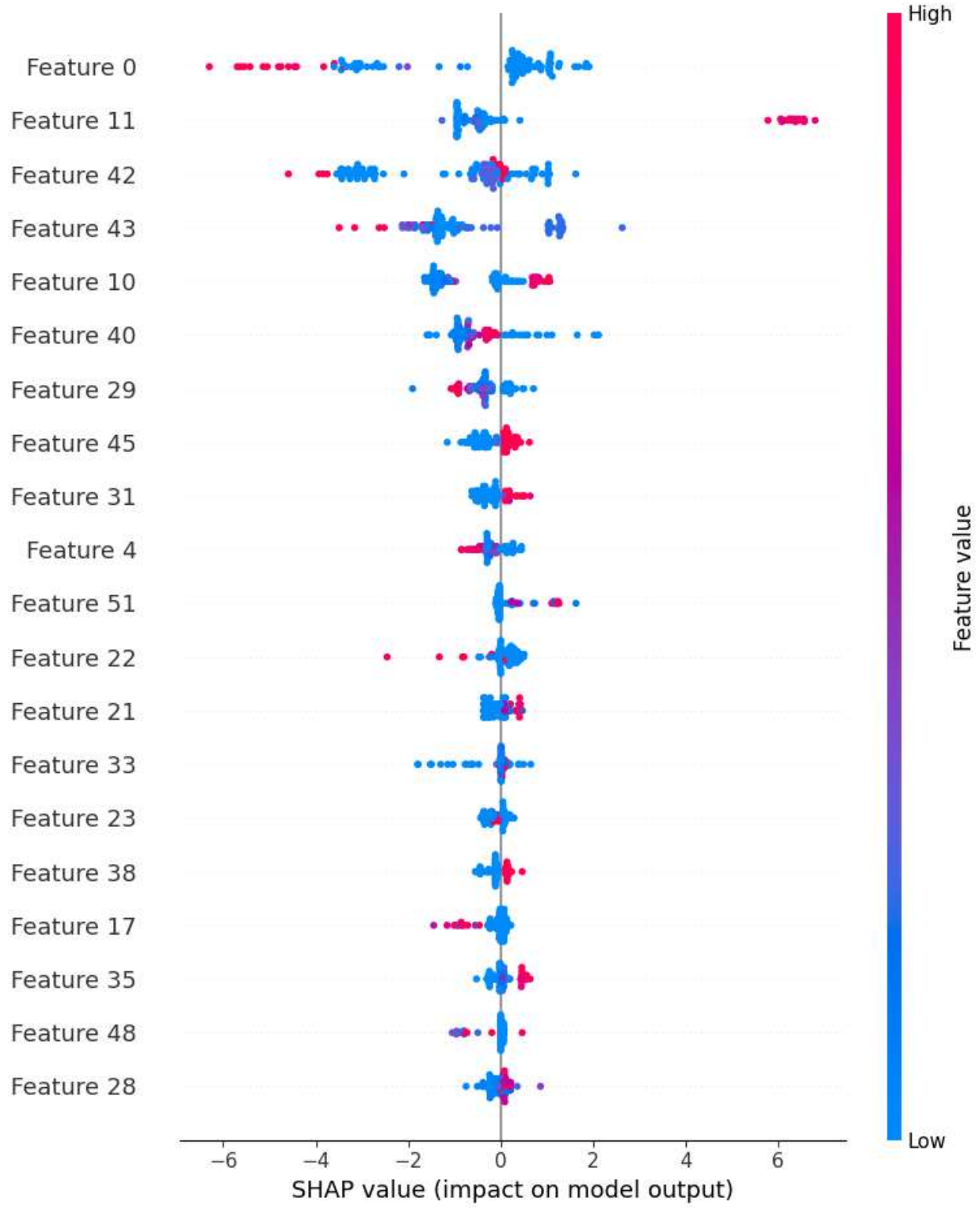


The ROC curve illustrates the model's ability to distinguish between benign and malicious traffic. The near-perfect ROC-AUC value confirms the strong discriminative capability of the model.

4.5 Explain ability Result

10.48047/jocaaa.2024.33.08.455

Fig 4: SHAP summary plot illustrating the impact of individual features on model predictions. The plot provides both the magnitude and direction of feature influence, enhancing interpretability and supporting explainable intrusion detection.



The SHAP summary plot highlights the most influential features affecting model predictions. This enhances transparency and provides valuable insights into the decision-making process of the IDS.

5.Discussion

10.48047/jocaaa.2024.33.08.455

The results demonstrate that the proposed model achieves superior performance compared to traditional intrusion detection approaches. The integration of XGBoost with SHAP enables both high accuracy and interpretability, addressing a critical challenge in cybersecurity. The use of a cleaned dataset significantly improves model performance by reducing noise and redundancy. However, the near-perfect results highlight the importance of validating the model on additional datasets to ensure generalizability and robustness. The explain ability provided by SHAP enhances trust in the model by enabling cybersecurity analysts to understand the factors influencing predictions. This is particularly important for real-world deployment, where transparency is essential.

Conclusion and Future Work

This paper presents an explainable intrusion detection system that combines optimized machine learning with interpretability. The proposed XGBoost-based model achieves excellent performance on the CICIDS2017 dataset and provides meaningful insights through SHAP analysis. The proposed framework is suitable for real-time cyber threat intelligence applications and offers a practical solution for modern cybersecurity challenges. Future work will focus on extending the model to multi-class classification, evaluating it on additional datasets, and deploying it in real-time environments.

References

1. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications, 60*, 19–31.
2. Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Network and Computer Applications, 25*, 152–160.
3. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
4. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176.
5. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
7. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
8. Dua, D., & Graff, C. (2019). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences.
9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
10. Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system. *MILCOM 2016–2016 IEEE Military Communications Conference*, 21–26.
11. Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications, 41*(4), 1690–1700.
12. Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., & Das, K. (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. *Proceedings DARPA Information Survivability Conference and Exposition (DISCEX'00), 2*, 12–26.
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

10.48047/jocaaa.2024.33.08.455

14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

15. Saxe, J., & Berlin, K. (2015). Deep neural network based malware detection using two dimensional binary program features. *arXiv preprint arXiv:1508.03096.*

16. Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 108–116.

17. Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence, 2*(1), 41–50.

18. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *2010 IEEE Symposium on Security and Privacy*, 305–316.

19. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–6.

20. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019). Robust intelligent intrusion detection systems using deep learning. *The Journal of Supercomputing, 75*, 3234–3248.

21. Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access, 5*, 21954–21961.

22. Zhang, J., Zulkernine, M., & Haque, A. (2008). Random-forest-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38*(5), 649–659.