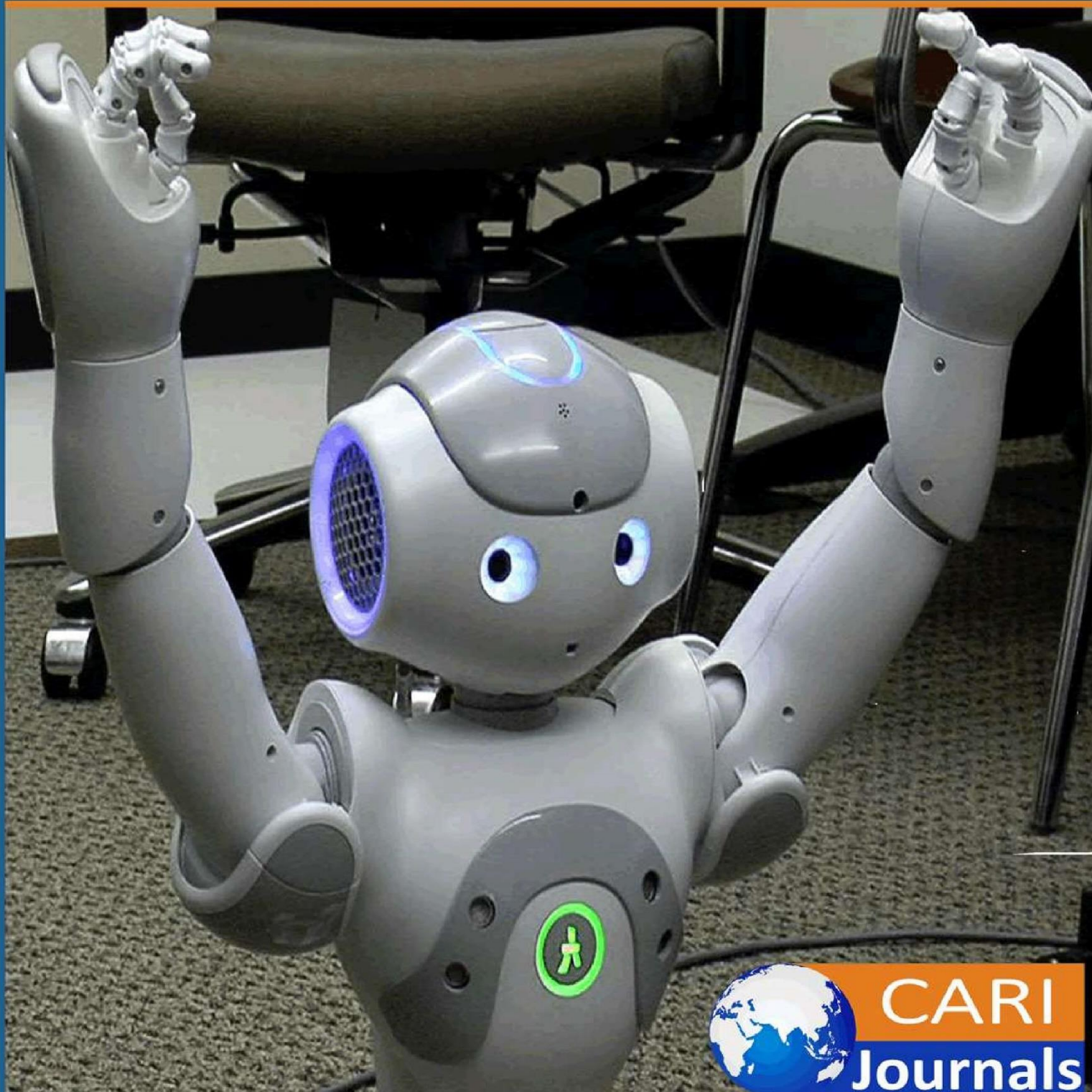
International Journal of
Computing and
Engineering
(IJCE)

Machine Learning for Intrusion Detection in Cloud-Based Systems



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 1, Issue No. 2, pp. 11 - 19, 2020
www.carijournals.org

Machine Learning for Intrusion Detection in Cloud-Based Systems

Tirumala Ashish Kumar Manne

Institution of Affiliation: Optum

https://orcid.org/0009-0009-9281-2930

Accepted: 16th May, 2020, Received in Revised Form: 1st Jun, 2020, Published: 16th Jun, 2020

Abstract

The proliferation of cloud computing has transformed data storage and processing but also introduced complex security challenges. Traditional Intrusion Detection Systems (IDS) often struggle in dynamic cloud environments due to scalability, adaptability, and the high rate of false positives. Machine Learning (ML) has emerged as a powerful tool to enhance IDS by enabling systems to learn from vast datasets, identify anomalous behavior, and adapt to evolving threats. This paper investigates the application of ML techniques such as supervised, unsupervised, and deep learning to intrusion detection in cloud-based systems. It reviews key methodologies, evaluates performance across widely used benchmark datasets (NSL-KDD, CICIDS2017), and highlights real-world implementations in commercial cloud platforms. The study also addresses critical challenges including data privacy, adversarial ML, real-time detection, and scalability. Through a comprehensive analysis, we identify promising research directions such as federated learning, explainable AI, and hybrid cloud-edge IDS architectures.

Keywords: *Intrusion Detection System (IDS), Machine Learning (ML), Cloud Security, Cybersecurity, Cloud Computing*

1. INTRODUCTION

Cloud computing has revolutionized the delivery of IT services by providing scalable, on-demand access to shared resources. Its dynamic, multi-tenant nature poses significant security challenges, particularly in detecting and mitigating cyber threats [1]. Intrusion Detection Systems (IDS) serve as a critical line of defense but often fall short in cloud environments due to limited scalability and adaptability [2]. Traditional IDS methods, such as signature-based and rule-based systems, struggle to identify novel or evolving attacks, resulting in high false positive rates and missed threats [3].

Machine Learning (ML) has emerged as a promising approach for enhancing IDS by enabling systems to learn patterns of normal and anomalous behavior from data. ML techniques, including supervised and unsupervised learning, can identify complex intrusion patterns, adapt to new threats, and reduce reliance on predefined signatures [4]. Deep learning architectures have shown improved detection accuracy in high-dimensional cloud data environments [5]. This paper explores the integration of ML in IDS tailored for cloud-based systems. It presents a comprehensive review of current techniques, evaluates existing models using benchmark datasets, and identifies critical challenges and future directions for research in this domain.

2. INTRUSION DETECTION IN CLOUD ENVIRONMENTS

Cloud computing environments offer scalable and flexible infrastructure but introduce a complex and dynamic attack surface. As organizations increasingly migrate their services to cloud platforms, the need for robust and intelligent security mechanisms, particularly Intrusion Detection Systems (IDS), has become paramount.

Types of Intrusions in Cloud-Based Systems:

Intrusions in cloud environments fall into several categories. External intrusions typically originate from attackers exploiting public interfaces and include threats such as Distributed Denial of Service (DDoS), phishing, and malware injection [6]. Internal intrusions stem from malicious insiders or compromised tenant accounts that misuse their privileges [7]. Data-focused intrusions, such as data leakage or unauthorized access, often result from misconfigurations or insecure APIs [8]. The complexity of multi-tenant architectures makes distinguishing legitimate access from malicious activity especially challenging.

Conventional IDS: Signature-Based vs. Anomaly-Based

Traditional IDS approaches are generally categorized into signature-based and anomaly-based systems. Signature-based IDS, like Snort, detect threats by comparing network traffic to known attack patterns [9]. While highly accurate for known threats, they are ineffective against zero-day exploits and evolving attack strategies. Anomaly-based IDS models baseline "normal" system behavior and flag deviations as potential intrusions. Although they can detect novel attacks, these systems are prone to high false positive rates, especially in dynamic cloud environments [10].

Limitations of Traditional IDS in the Cloud:

Traditional IDS frameworks, originally designed for static, on-premises networks, face limitations when deployed in cloud environments. The elasticity and virtualization features of cloud platforms render conventional traffic-monitoring strategies less effective [11]. Additionally, the lack of visibility into lower-level infrastructure due to abstraction layers introduced by cloud service providers (CSPs) complicates deployment and reduces detection accuracy [12]. Traditional IDS solutions are typically not scalable, making them ill-suited for handling the volume and velocity of cloud traffic.

3. Machine Learning Techniques for IDS

Machine Learning (ML) techniques have emerged as powerful tools in designing Intrusion Detection Systems (IDS) for cloud environments due to their ability to identify complex patterns, detect zero-day attacks, and adapt to evolving threats. This section outlines the primary categories of ML used in IDS: supervised, unsupervised, semi-supervised, and deep learning approaches.

Supervised Learning

Supervised learning involves training models on labeled datasets to classify network behavior as normal or malicious. Algorithms like Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes (NB), and Random Forests (RF) have been extensively used in IDS research [13]. These models provide high accuracy when trained with quality data but struggle with generalization to novel attacks. For example, SVMs have demonstrated strong performance on benchmark datasets like NSL-KDD, yet require careful feature selection and parameter tuning [14].

Unsupervised Learning

Unsupervised methods identify anomalies without requiring labeled data, making them suitable for detecting unknown threats. Clustering techniques such as K-Means, DBSCAN, and hierarchical clustering are common [15]. These approaches are useful in dynamic cloud environments, but they may suffer from high false positives due to overlapping behavior between benign and malicious patterns [16].

Semi-Supervised Learning

Semi-supervised techniques leverage a small amount of labeled data along with a larger pool of unlabeled data, offering a practical solution when labeled intrusion data is scarce. Methods such as self-training and co-training enhance model generalizability while reducing the cost of data annotation [17].

Deep Learning Approaches

Deep learning (DL) models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, have shown superior performance in learning hierarchical representations from raw data [18]. For instance, LSTM models can detect complex sequential intrusion patterns in network flows, making them

particularly effective for time-series traffic analysis in cloud systems [19]. Autoencoders have also been applied for unsupervised anomaly detection, offering scalability and adaptability [20].

Reinforcement Learning for Adaptive IDS

Reinforcement Learning (RL) is gaining traction in adaptive IDS design, where agents learn optimal policies for intrusion response through interaction with the environment. Techniques such as Q-learning have been explored to dynamically update detection policies in real-time cloud settings [21].

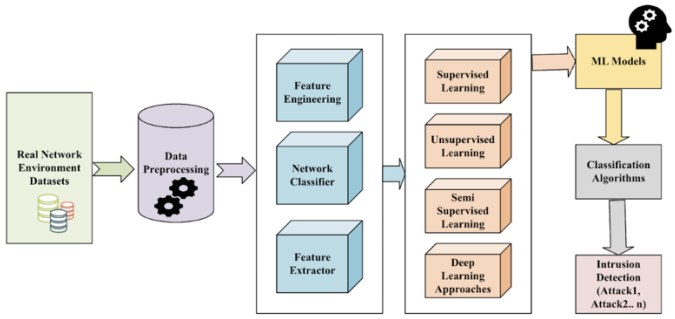


Figure 1. Machine Learning for Intrusion Detection System

4. Datasets and Evaluation Metrics

A critical aspect of developing and evaluating Machine Learning (ML)-based Intrusion Detection Systems (IDS) is the availability of representative datasets and appropriate performance metrics. The quality and relevance of training data directly influence a model's ability to generalize and detect novel threats in cloud environments.

Common Datasets for IDS Research

Several benchmark datasets have been widely used in the literature for training and evaluating IDS models.

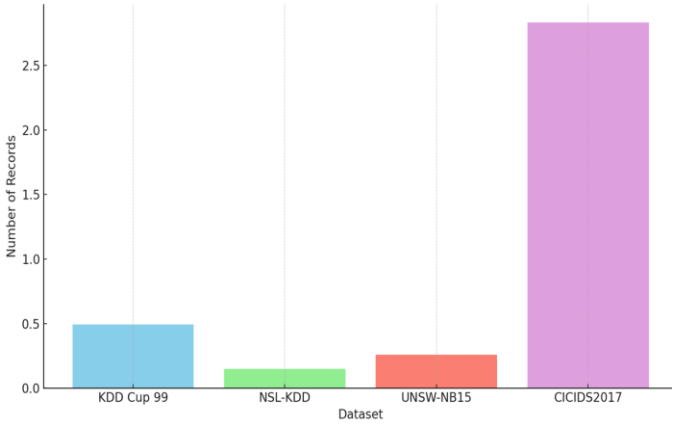


Figure 1. Number Of Records in IDS Datasets

KDD Cup 99: One of the earliest datasets, derived from DARPA 1998, it includes simulated attacks categorized into four types: DoS, U2R, R2L, and probing. Despite its popularity, it has been criticized for redundant records and outdated attack profiles [22].

NSL-KDD: A refined version of KDD Cup 99, NSL-KDD addresses issues such as duplicate records and class imbalance. It is widely used to benchmark both supervised and unsupervised ML techniques [23].

UNSW-NB15: Developed to overcome the limitations of KDD-based datasets, it includes modern attack vectors and realistic traffic captured using the IXIA PerfectStorm tool. It provides a balanced distribution of normal and attack data [24].

CICIDS2017: Created by the Canadian Institute for Cybersecurity, this dataset includes recent attack scenarios such as brute-force, botnet, and DDoS, captured in realistic network environments [25].

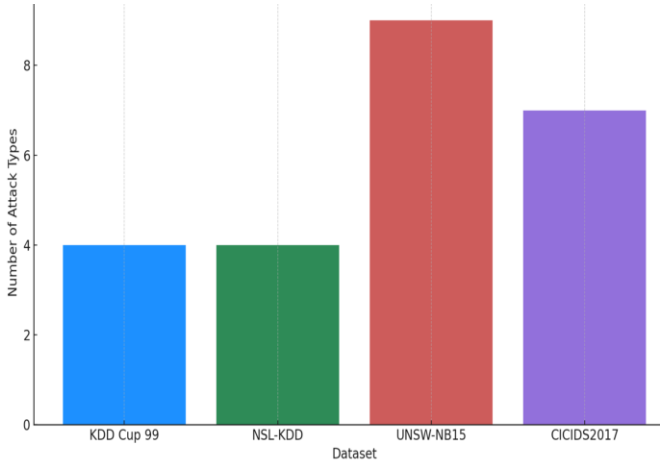


Figure 2. Number Of Attack Types in IDS Datasets

Evaluation Metrics

Evaluating the performance of IDS models requires appropriate metrics, especially in the presence of class imbalance and varying attack types.

Accuracy: Measures the proportion of correctly classified instances. However, it can be misleading in imbalanced datasets [26].

Precision and Recall: Precision quantifies the number of true positives among all predicted positives, while recall (or true positive rate) indicates how many actual attacks are correctly identified. These are particularly important for anomaly detection [27].

F1-Score: The harmonic mean of precision and recall, offering a balance between them in skewed datasets [28].

Receiver Operating Characteristic (ROC) and Area Under Curve (AUC): These are used to visualize the trade-off between true positive and false positive rates. AUC provides a single measure of model performance regardless of threshold [29].

In cloud environments, real-time performance, false positive rate, and scalability are also key considerations, as excessive alerts can overwhelm security teams and degrade trust in the system.

5. Potential Uses

Industry Benchmarking: Security professionals can use the insights to benchmark their existing IDS systems and evaluate the feasibility of adopting ML-based alternatives.

Cloud Security Framework Design: Architects designing secure cloud platforms can use the article's analysis of intrusion types and datasets to inform IDS framework decisions.

Policy and Governance Guidance: Government or regulatory bodies may reference the article when setting standards or policies for intrusion detection and response in cloud environments.

Threat Intelligence Enhancement: SOC (Security Operations Center) teams can incorporate the article's taxonomy of attacks and ML-based detection approaches to improve threat response strategies.

Tool Evaluation Criteria: Organizations evaluating IDS tools can use the dataset and metric comparison charts as part of their vendor assessment process.

6. Conclusion

As cloud computing continues to transform the digital landscape, securing cloud-based systems against evolving cyber threats remains a top priority. Traditional intrusion detection systems, while useful, are increasingly inadequate in handling the dynamic, high-volume, and distributed nature of cloud environments. This article has examined the potential of machine learning techniques to enhance intrusion detection through improved adaptability, precision, and scalability. We reviewed various ML approaches including supervised, unsupervised, semi-supervised, and deep learning models highlighting their respective strengths and challenges. The analysis of widely-used datasets such as NSL-KDD, UNSW-NB15, and CICIDS2017, alongside evaluation metrics like accuracy, precision, recall, and AUC, provides a clear foundation for benchmarking IDS performance.

Despite promising advancements, several challenges remain, including dataset relevance, high false positive rates, data privacy, and the threat of adversarial attacks. Future research should focus on developing explainable, federated, and context-aware IDS models that align with real-time, multi-tenant cloud environments. Machine learning offers a powerful paradigm for intrusion detection, but its effective integration into cloud security frameworks demands continuous innovation, interdisciplinary collaboration, and rigorous validation.

REFERENCES

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," NIST Special Publication 800-145, 2011.

[2] M. A. Baig, "A Systematic Review of Cloud Security Challenges in Cloud Computing," J. Cloud Comput., vol. 6, no. 1, 2017.

[3] S. Xie et al., "Anomaly detection in cloud computing using machine learning: A review," IEEE Access, vol. 7, pp. 177421–177433, 2019.

[4] A. Javaid et al., "A deep learning approach for network intrusion detection system," Proc. 9th EAI Int. Conf. Bio-inspired Info. Commun. Technol., 2016.

[5] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), 2015.

[6] S. Roschke, F. Cheng, and C. Meinel, "Intrusion detection in the cloud," 8th IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009, pp. 729–734.

[7] M. H. Sqalli, F. Al-Haidari, and K. Salah, "EDoS-Shield - A two-steps mitigation technique against EDoS attacks in cloud computing," 8th IEEE International Conference on Computer Engineering & Systems, 2012.

[8] A. Khorshed, A. Ali, and S. Wasimi, "A survey on gaps, threats and attacks in cloud computing," Journal of Internet Services and Applications, vol. 4, no. 1, pp. 1–9, Jan. 2013.

[9] M. Roesch, "Snort: Lightweight intrusion detection for networks," Proceedings of the 13th USENIX conference on System administration, 1999, pp. 229–238.

[10] R. Mitchell and I. R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," ACM Computing Surveys (CSUR), vol. 46, no. 4, pp. 1–29, Mar. 2014.

[11] H. Gonzalez et al., "Cloud security auditing: Challenges and emerging approaches," IEEE Security & Privacy, vol. 10, no. 5, pp. 12–19, Sept.-Oct. 2012.

[12] J. Zhang and B. H. Kang, "An overview of intrusion detection in cloud computing," Proceedings of the 2013 International Conference on IT Convergence and Security (ICITCS), 2013, pp. 1–4.

[13] M. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," IEEE Transactions on Computers, vol. 65, no. 10, pp. 2986–2998, Oct. 2016.

[14] N. Moustafa and J. Slay, "The significant features of the UNSW-NB15 dataset for network intrusion detection systems," Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), 2015.

[15] F. A. A. Elrahman and A. Abraham, "A review of class imbalance problem in intrusion detection," Journal of Network and Computer Applications, vol. 75, pp. 35–54, Nov. 2016.

[16] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Networks, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.

[17] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, "HIDE: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification," Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, 2001.

[18] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with deep learning," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 43–48, 2017.

[19] S. Kim, K. Lee, and H. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, Mar. 2014.

[20] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), 2016.

[21] Y. Sun, Z. Zhang, and Y. Li, "Reinforcement learning-based adaptive system for network intrusion detection," Journal of Electrical and Computer Engineering, vol. 2014, Article ID 139275, 2014.

[22] S. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD Cup 99 data set," Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[23] M. Mahoney and P. Chan, "An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection," RAID 2003: Recent Advances in Intrusion Detection, Springer, pp. 220–237, 2003.

[24] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Communications and Information Systems Conference (MilCIS), IEEE, 2015.

[25] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," ICISSP, pp. 108–116, 2018.

[26] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, June 2006.

[27] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[28] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 233–240, 2006.

[29] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," Pattern Recognition Letters, vol. 30, no. 1, pp. 27–38, Jan. 2009.