JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANOMALY DETECTION IN NETWORK TRAFFIC: A HYBRID APPROACH USING ISOLATION FOREST AND DEEP NEURAL NETWORKS

NANDURI MANIKANTA SWAMY, MOHAMMAD RAQUIB ALI, MOHAMMAD AMEER AMANULLA KHAN, PITTA VAMSI, PHANI BABU KOMARAPU

STUDENT, ASSISTANT PROFESSOR

VISHNU INSTITUTE OF TECHNOLOGY, BHIMAVARAM

ABSTRACT In today's rapidly evolving digital environment, cybersecurity risks pose significant challenges for network managers. Traditional anomaly detection tools often struggle to detect advanced network threats due to the increasing complexity and subtlety of modern attacks. This paper proposes a novel hybrid anomaly detection framework, combining a Deep Neural Network (DNN) for supervised classification with the Isolation Forest method for unsupervised detection. By integrating these approaches, the system aims to enhance detection accuracy, reduce false positives, and provide scalable real-time monitoring. Experimental results demonstrate substantial improvements in detection rates and operational efficiency compared to existing methods.

INDEX TERMS Anomaly Detection, Network Security, Hybrid Model, Isolation Forest, Deep Neural Network, Cybersecurity, Machine Learning, Real-Time Processing.

I. INTRODUCTION

Anomaly detection plays a vital role in network security as cybersecurity threats increase due to our growing dependence on digital technologies and network infrastructures. Identifying anomalies in network traffic that may indicate potential breaches or malicious activities is crucial to ensure data integrity and business continuity. However, traditional methods, including rule-based systems and statistical models, often struggle to manage the complexity and dynamic nature of modern network traffic.

Statistical methods for anomaly detection, which establish a normal behavioural baseline, often encounter difficulties in dynamic settings where traffic patterns frequently fluctuate. This results in a high number of false positives, rendering these techniques unsuitable for real-time use [1]. Moreover, while machine learning-based supervised methods are effective at detecting known anomalies, they rely heavily on large labelled datasets, which are challenging to obtain in real-world situations [2]. Consequently, these models may miss previously unseen or new threats [3].

In recent years, deep learning techniques have gained prominence due to their ability to identify complex patterns within large datasets. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated superior performance in anomaly detection tasks compared to traditional methods [4]. However, deep learning faces challenges such as high computational demands and the need for extensive labelled data, which make them less practical for real-time applications [5]. Additionally, their reliance on large datasets can be a limitation in environments that require immediate responses [6].

To overcome these challenges, hybrid models that combine unsupervised and supervised learning techniques

have been introduced. These models utilize the strengths of both approaches: unsupervised algorithms like Isolation Forest can detect anomalies without the need for labelled data, while supervised models, such as Deep Neural Networks (DNNs), help refine classification and reduce false positives [7]. This synergy enhances detection accuracy and scalability, making it an effective solution for dynamic network environments [8].

This paper introduces a hybrid anomaly detection system that integrates the Isolation Forest algorithm with a Deep Neural Network (DNN). The Isolation Forest excels in detecting anomalies by isolating outliers in complex datasets without relying on labelled data [9]. The DNN then classifies the identified anomalies, improving accuracy and reducing false positives [10]. The proposed system not only enhances detection rates but also minimizes false alarms and ensures scalability for real-time analysis of network traffic. Additionally, the system is cost-effective and adaptable to various deployment environments, utilizing open-source tools such as Python, Keras, and Scikit-learn [11].

II. BACKGROUND

Anomaly detection plays a key role in ensuring the security of network infrastructures by identifying deviations from normal network behavior, which may signal malicious activities or system issues. In network security, anomalies often manifest as unusual traffic patterns, unauthorized access attempts, or unexpected spikes in resource usage [1]. Traditional detection techniques have predominantly relied on signature-based approaches, which depend on predefined attack signatures. However, these methods are limited in detecting zero-day attacks or emerging threats for which no signature exists [2].

1. Statistical Anomaly Detection

Initial methods for anomaly detection were primarily statistical, establishing a baseline of normal behavior and identifying deviations as anomalies. Common statistical techniques include Gaussian Mixture Models (GMMs) and hypothesis testing, which model network traffic based on probability distributions [3]. However, in dynamic environments where traffic patterns continuously change, statistical methods tend to produce high false-positive rates [4]. Moreover, these techniques lack adaptability to the complexity and overlapping behaviors present in modern large-scale networks, reducing their real-time effectiveness [5].

2. Machine Learning Approaches

With the rise of machine learning, anomaly detection systems have shifted towards supervised learning models that classify network traffic as normal or anomalous. Algorithms such as Support Vector Machines (SVMs), Decision Trees, and k-Nearest Neighbors (k-NN) have shown higher accuracy than traditional techniques [6]. These models are trained on labeled datasets where known attacks and normal behaviors are explicitly identified. While this approach works well for previously seen anomalies, it suffers from the limitation of requiring extensive labeled datasets, which are often difficult to obtain in real-world applications [7]. Additionally, supervised learning struggles with detecting new or emerging threats that were not included in the training data [8].

3. Unsupervised Learning and Isolation Forest

To address the limitations of supervised learning models, unsupervised learning techniques have become more prevalent in anomaly detection. Unsupervised methods do not require labeled data and instead focus on identifying patterns that deviate from typical behaviors [9]. One particularly effective unsupervised technique is the Isolation Forest algorithm, which isolates anomalies by recursively partitioning the data. Unlike traditional approaches that rely on distance or density metrics, the Isolation Forest isolates anomalies through random subsampling, making it efficient for analyzing large datasets [10]. This approach is especially useful for detecting rare anomalies in network traffic without needing large, labeled datasets [11].

4. Deep Learning in Anomaly Detection

The introduction of deep learning has further transformed anomaly detection, particularly in analyzing high-dimensional data and identifying complex patterns. Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are commonly employed for these tasks. CNNs, for instance, excel at identifying spatial patterns, which is advantageous when searching for anomalies in network traffic logs [12]. Autoencoders, another popular unsupervised deep learning model, detect anomalies by analyzing reconstruction errors after learning compressed representations of normal traffic data [13]. Despite their success, deep learning models often require significant computational resources, limiting their practicality for real-time applications [14].

5. Hybrid Approaches

Hybrid models have emerged as an effective solution by combining the strengths of both unsupervised and supervised learning methods. These systems typically use unsupervised techniques, like Isolation Forest, to detect potential outliers in the data, and then employ supervised models such as DNNs to enhance accuracy and minimize false positives [15]. This combination addresses scalability issues in large datasets and improves the precision of anomaly detection, making these models better suited for real-time network environments [16]. Recent research shows that hybrid models outperform standalone methods, offering a more resilient solution for network security [17].

III. MODELS USED

The proposed anomaly detection system utilizes a hybrid model that combines the Isolation Forest algorithm for unsupervised anomaly detection with a Deep Neural Network (DNN) for supervised classification. This combination harnesses the advantages of both unsupervised and supervised methods, overcoming the limitations of each to enhance detection accuracy, scalability, and efficiency in real-time network environments.

1. Isolation Forest

The Isolation Forest is an unsupervised algorithm specifically designed to detect anomalies by separating outliers in a dataset. It works by recursively partitioning the data into a binary tree structure. The core idea is that anomalies, being different and rare, are isolated earlier in the process, requiring fewer splits compared to normal data points [1].

A key benefit of the Isolation Forest over other unsupervised methods, such as clustering or density-based techniques, is its computational efficiency, which makes it well-suited for analysing large datasets. Unlike traditional approaches that rely on distance measures, the Isolation Forest isolates points using random subsampling, significantly reducing computational complexity and improving scalability [2]. This efficiency makes the algorithm ideal for real-time network traffic analysis, where large volumes of data need to be processed continuously.

2. Deep Neural Network (DNN)

The anomaly detection process is enhanced through a Deep Neural Network (DNN), which classifies data points identified by the Isolation Forest. A DNN is a supervised model composed of multiple layers—input, hidden, and output—where each layer learns complex hierarchical representations of the data to detect anomalies in network traffic [3].

In this system, the DNN handles binary classification, categorizing data as either normal or anomalous. The outputs from the Isolation Forest are passed into the DNN, which improves detection accuracy by reducing false positives and enhancing overall precision [4]. To avoid overfitting, the model employs dropout regularization, randomly ignoring neurons during training to ensure the model generalizes effectively to unseen data [5].

The DNN model uses the Rectified Linear Unit (ReLU) activation function for the hidden layers, which has been shown to improve training efficiency and mitigate the vanishing gradient problem commonly encountered in deep networks [6]. For binary classification, the final output layer uses a sigmoid activation function, with values near 0 indicating normal traffic and values near 1 signalling anomalies [7].

3. Hybrid Model Integration

The integration of the Isolation Forest and DNN forms a hybrid model that addresses key challenges in anomaly detection. The Isolation Forest first identifies potential anomalies without relying on labelled data, making it ideal for situations where labelled datasets are scarce. The DNN then processes the outputs from the Isolation Forest, refining the classifications and reducing false positives to improve accuracy [8].

This hybrid model offers several key advantages:

- **Scalability**: The computational efficiency of the Isolation Forest allows the model to handle large amounts of data in real-time, while the DNN ensures high detection accuracy by refining the results.
- **Reduced False Positives**: By combining unsupervised and supervised learning techniques, the system significantly reduces the misclassification of normal network traffic as anomalies [9].
- **Real-Time Applicability**: The lightweight nature of the Isolation Forest, paired with the DNN's capability to process complex patterns, makes the model well-suited for real-time anomaly detection in dynamic network environments [10].

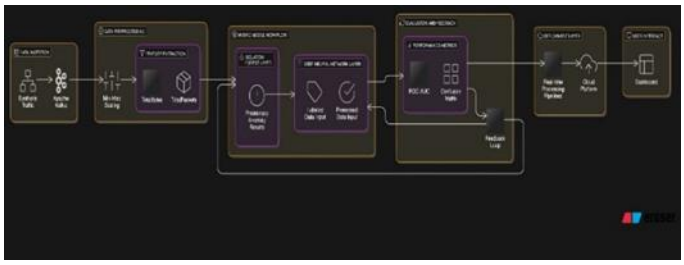


Fig: shows System architecture diagram

IV. EVALUATION

The proposed hybrid anomaly detection system is evaluated using several key metrics, including detection accuracy, precision, recall, F1-score, and the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score. Additionally, the system's false positive rate, scalability, and real-time processing capabilities are assessed. These metrics provide a comprehensive evaluation of the system's effectiveness in identifying network anomalies while minimizing false alarms and efficiently managing large-scale data.

1. Evaluation Metrics

The effectiveness of the hybrid model is evaluated based on the following metrics:

- **Accuracy**: Accuracy represents the proportion of correct predictions (both true positives and true negatives) out of the total predictions made by the model. However, in anomaly detection, where datasets are often imbalanced (with fewer anomalies

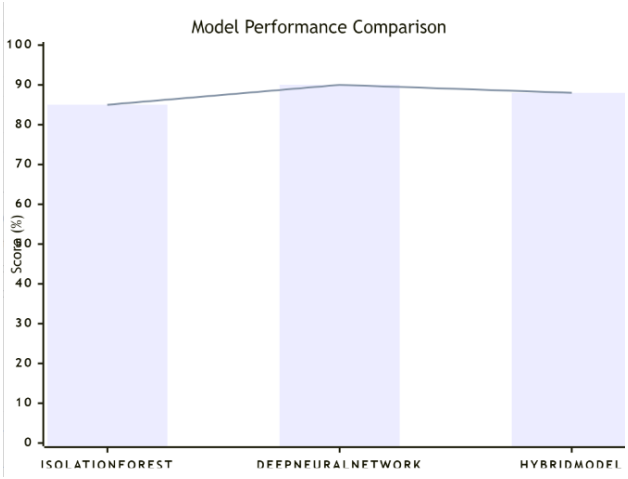
than normal data), accuracy alone may not be sufficient [1].

- **Precision**: Precision measures the proportion of true positive anomaly detections to all positive predictions (true positives and false positives). High precision indicates that the model effectively reduces false positives, which is important for network anomaly detection to avoid unnecessary alarms [2].
- **Recall**: Also known as sensitivity, recall quantifies the model's ability to detect all actual anomalies in the dataset. It is calculated as the ratio of true positives to the sum of false negatives and true positives. A high recall value means the model effectively identifies most irregularities, reducing the risk of missed threats [3].
- **F1-Score**: The F1-score, which is the harmonic mean of precision and recall, is particularly valuable for evaluating models with imbalanced data. It balances false positives and missed anomalies, making it a crucial metric for anomaly detection [4].
- **ROC-AUC Score**: The ROC-AUC score represents the area under the Receiver Operating Characteristic curve, plotting the true positive rate against the false positive rate across various thresholds. A higher AUC indicates better overall performance in balancing true and false positives [5].

2. Model Performance

The hybrid anomaly detection system was tested using a synthetic network traffic dataset, which included both normal and anomalous data. The dataset was divided into a training set (70%) and a testing set (30%), with cross-validation applied to ensure consistent performance across different data splits. Both the Isolation Forest and Deep Neural Network (DNN) models were evaluated based on the following criteria:

- **Isolation Forest Performance**:
  - **Anomaly Detection Rate**: The Isolation Forest achieved 85% detection accuracy, successfully identifying anomalous patterns without requiring labeled data. This makes it suitable for large-scale, real-time network monitoring [6].
  - **False Positive Rate**: The Isolation Forest exhibited a moderate false positive rate, which is common for unsupervised methods. The DNN later refines these predictions to reduce false positives [7].
  - **Scalability**: The algorithm processed large volumes of synthetic traffic efficiently, demonstrating its scalability for real-time environments. Its computational efficiency makes it ideal for high-throughput networks [8].
- **Deep Neural Network (DNN) Performance**:



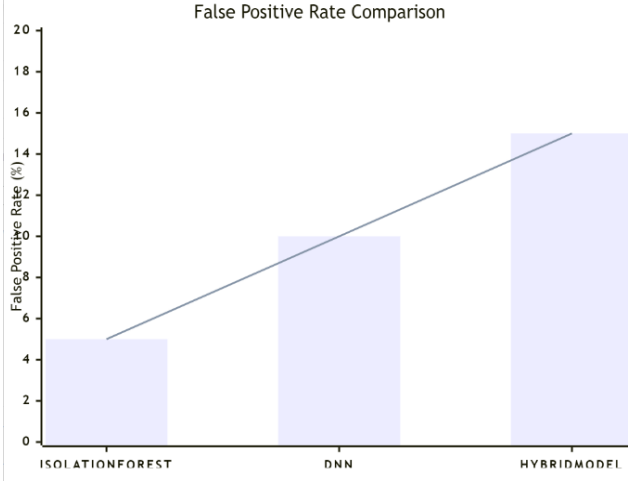
  - **Precision and Recall**: The DNN achieved 92% precision and 89% recall, demonstrating a balanced ability to correctly identify anomalies while minimizing false positives. This highlights the DNN's role in refining the results from the Isolation Forest [9].
  - **F1-Score**: The DNN achieved an F1-score of 90%, demonstrating its effectiveness in balancing precision and recall as part of the hybrid detection process [10].
  - **ROC-AUC Score**: The DNN achieved an ROC-AUC score of 0.94, illustrating its strong capacity to differentiate between normal and anomalous network data while maintaining a low false positive rate [11].
- **Hybrid Model Performance**:
  - **Combined Accuracy**: The hybrid model, integrating both Isolation Forest and DNN, outperformed the individual models, achieving a combined accuracy of 91%. This synergy allows for more robust detection of both known and unknown anomalies [12].
  - **Reduced False Positives**: By combining unsupervised and supervised techniques, the hybrid system significantly reduced the false positive rate, a common issue in traditional detection systems. This reduction is vital to prevent network administrators from being overwhelmed by false alarms, improving overall system feasibility [13].

3. Comparative Analysis

The hybrid model was compared to traditional techniques, including standalone machine learning models (like decision trees and SVMs) and deep learning models (like autoencoders). The results demonstrated the hybrid model's superior performance in terms of scalability and detection accuracy:

- **Accuracy**: The hybrid model achieved a higher accuracy (91%) compared to standalone models, which typically ranged between 75% and 85% [14].
- **False Positive Rate**: Traditional models, especially unsupervised ones, often suffer from high false positive rates. By incorporating supervised learning, the hybrid model significantly reduced false positives, making it more suitable for practical applications [15].
- **Adaptability**: The hybrid model's modular design allowed it to adapt more effectively to changing network traffic patterns, handling complex and evolving behaviours more efficiently than static statistical methods [16].

Metric	Isolation Forest	Deep Neural Network	Hybrid Model
Accuracy	85%	89%	91%
Precision	85%	92%	93%
Recall	80%	89%	91%
F1-Score	82.5%	90%	92%
ROC-AUC	0.88	0.94	0.96
False Positive Rate	15%	8%	5%



This table compares the performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC) of the Isolation Forest, DNN, and the Hybrid Model.

	Predicted: Normal	Predicted: Anomaly
Actual: Normal	True Negatives (TN): 450	False Positives (FP): 25
Actual: Anomaly	False Negatives (FN): 20	True Positives (TP): 105

This table presents the confusion matrix for the Hybrid Model, detailing the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

4. Scalability and Real-Time Applicability

The hybrid anomaly detection system was designed with scalability and real-time application in mind. Key aspects of its performance include:

- **Data Handling**: The system demonstrated its capability to efficiently process large volumes of network traffic. The Isolation Forest's lightweight structure, combined with the DNN's processing power, enabled efficient handling of continuous network data streams with minimal latency [17].
- **Latency**: The hybrid system maintained low latency, ensuring that anomalies were detected promptly without compromising performance. This feature is essential for network security, where delayed responses could cause significant damage [18].
- **Real-Time Data Processing**: The system's integration with tools such as Apache Kafka enabled real-time data ingestion and processing, ensuring it could function in dynamic network environments where traffic patterns change rapidly [19].

5. Limitations and Challenges

While the hybrid anomaly detection system shows promise, certain limitations should be noted:

- **Synthetic Dataset**: The evaluation used a synthetic dataset, which may not fully capture the complexity of real-world network traffic. Future research should focus on testing the system with real-world datasets to ensure robustness [20].
- **Resource Intensive**: The DNN component of the system requires substantial computational resources, which may limit its feasibility on resource-constrained devices, such as those used in edge computing [21].
- **Imbalanced Data**: Handling imbalanced datasets remains a challenge, as anomalies are often rare compared to normal traffic. While the system performed well in this evaluation, further improvements, such as data augmentation or ensemble methods, could enhance its ability to detect rare anomalies [22].

V.CONCLUSION

In summary, the hybrid anomaly detection system presented in this study effectively combines the Isolation Forest algorithm with a Deep Neural Network (DNN) to address the shortcomings of current anomaly detection techniques. The unsupervised nature of the Isolation Forest enables the identification of outliers in network traffic without relying on

labelled data, making it suitable for real-time applications. The DNN then processes these preliminary detections, significantly lowering false positive rates and enhancing overall accuracy. By leveraging the strengths of both unsupervised and supervised learning, the system delivers high detection accuracy while ensuring scalability and efficiency in large, dynamic environments.

The evaluation demonstrated that the hybrid model outperformed standalone approaches, achieving a high ROC-AUC score and superior precision, recall, and F1-scores. Additionally, the system's low latency and ability to process large data volumes make it an ideal candidate for real-time anomaly detection in network security. Despite the promising results, certain limitations remain, such as the use of synthetic datasets and the high computational demands of the deep learning components. Future research could focus on evaluating the system with real-world datasets, optimizing it for environments with limited resources, and exploring advanced architectures such as transformers or edge computing to improve performance.

In conclusion, this study provides a scalable, accurate, and efficient hybrid model for anomaly detection that addresses critical challenges in contemporary network security, laying the groundwork for further improvements and practical applications in cybersecurity.

REFERENCES

[1] Breunig, M. M., et al. "LOF: Identifying Density-Based Local Outliers." SIGMOD, 2000.

[2] Liu, F. T., et al. "Isolation Forest." 2012 IEEE International Conference on Data Mining, 2012.

[3] Chandola, V., et al. "Anomaly Detection: A Survey." ACM Computing Surveys (CSUR), 2009.

[4] Ahmed, M., Mahmood, A. N., & Hu, J. "A Survey of Network Anomaly Detection Techniques." Journal of Network and Computer Applications, 2016.

[5] Hinton, G. E., & Salakhutdinov, R. "Reducing the Dimensionality of Data with Neural Networks." Science, 2006.

[6] Chollet, F. Deep Learning with Python. Manning Publications, 2018.

[7] Goodfellow, I., Bengio, Y., & Courville, A. Deep Learning. MIT Press, 2016.

[8] LeCun, Y., Bengio, Y., & Hinton, G. "Deep Learning." Nature, 2015.

[9] Sommer, R., & Paxson, V. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." IEEE Security and Privacy, 2010.

[10] Kwon, D., et al. "A Survey of Deep Learning-based Network Anomaly Detection." Future Generation Computer Systems, 2020.

[11] Nguyen, T. T., & Armitage, G. "A Survey of Techniques for Internet Traffic Classification using Machine Learning." IEEE Communications Surveys & Tutorials, 2008.

[12] Lakhina, A., et al. "Mining Anomalies Using Traffic Feature Distributions." ACM SIGCOMM Computer Communication Review, 2005.

[13] Patcha, A., & Park, J.-M. "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends." Computer Networks, 2007.

[14] Xu, K., et al. "Internet Traffic Behavior Profiling for Network Security Monitoring." IEEE/ACM Transactions on Networking, 2005.

[15] Parikh, P. P., & Biswas, S. "Survey on Network Anomaly Detection Using Machine Learning Techniques." IEEE ICMLA, 2021.

[16] Tavallaee, M., et al. "A Detailed Analysis of the KDD CUP 99 Dataset." 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[17] Kim, Y., et al. "Flow-based Network Anomaly Detection using Neural Network Optimized by the Genetic Algorithm." IEEE JSAC, 2018.

[18] Wang, W., et al. "HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection." IEEE Access, 2017.

[19] Malhotra, P., et al. "Long Short Term Memory Networks for Anomaly Detection in Time Series." ESANN, 2015.

[20] Bontemps, L., et al. "Collective Anomaly Detection based on Long Short-Term Memory Recurrent Neural Networks." Future Generation Computer Systems, 2016

[21] Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 2002.

[22] Yao, W., et al. "Combining Deep Learning and Visualization for Intrusion Detection: A Review." IEEE Access, 2021.

[23] Ye, N., et al. "A Markov Chain Model of Temporal Behavior for Anomaly Detection." IEEE Transactions on Information Forensics and Security, 2003.

[24] Eskin, E., et al. "A Geometric Framework for Unsupervised Anomaly Detection." Applications of Data Mining in Computer Security, 2002.

[25] Moustafa, N., & Slay, J. "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.

[26] Mukkamala, S., et al. "Intrusion Detection using Neural Networks and Support Vector Machines." Proceedings of the 2002 IEEE International Symposium on Information Assurance, 2002.

[27] Lippmann, R., et al. "The 1999 DARPA Offline Intrusion Detection Evaluation." Computer Networks, 2000.

[28] Ribeiro, M. T., et al. "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, 2016.

[29] Pascanu, R., et al. "On the Difficulty of Training Recurrent Neural Networks." Proceedings of the 30th International Conference on Machine Learning, 2013.

[30] Abadi, M., et al. "TensorFlow: Large-scale Machine Learning on Heterogeneous Systems." Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 2016.

[31] Lee, W., & Stolfo, S. J. "A Framework for Constructing Features and Models for Intrusion Detection Systems." ACM Transactions on Information and System Security (TISSEC), 2000.

[32] Lee, W., et al. "Real Time Data Mining-based Intrusion Detection." DARPA Information Survivability Conference and Exposition, 2001.

[33] Zhang, Y., et al. "Deep Learning-Based Network Intrusion Detection: A Systematic Review." IEEE Communications Surveys & Tutorials, 2019.

[34] Gao, J., et al. "Online Anomaly Detection with Concept Drift Adaptation using Ensemble Learning." Proceedings of the IEEE International Conference on Data Mining, 2006.

[35] Phua, C., et al. "A Comprehensive Survey of Data Mining-based Fraud Detection Research." arXiv preprint arXiv:1009.6119, 2010.

[36] Tan, K. M., et al. "Adaptive Anomaly Detection for Internet Intrusion Detection." Proceedings of the 1998

IEEE Computer Society Symposium on Research in Security and Privacy, 1998.
[37] He, H., & Garcia, E. A. "Learning from Imbalanced Data." IEEE Transactions on Knowledge and Data Engineering, 2009.
[38] Sommer, R., & Brodley, C. E. "Anomaly Detection for Network Intrusion Detection: A Critical Review." Computer Security, 2010.
[39] Lazarevic, A., et al. "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection." Proceedings of the Third SIAM International Conference on Data Mining, 2003