A Machine Learning-Based Approach with the Cyber Security Big Data (CSBD) System

Arpita Gupta [1], Dr. Shailja Sharma [2]

[1] *Research Scholar, Rabindranath Tagore University, Bhopal*
[2] *Profesor, Rabindranath Tagore University, Bhopal*

Abstract:- This paper presents a comprehensive analysis of various advanced intrusion detection systems (IDS) and their methodologies, emphasizing the integration of machine learning algorithms and feature selection techniques to enhance network security. It reviews several innovative approaches proposed in recent studies, each offering unique strategies to improve the accuracy and efficiency of intrusion detection in information technology networks. This paper introduces the Cyber Security Big Data (CSBD) system, a novel approach aimed at enhancing security within big data environments. The CSBD system serves as a platform for selecting appropriate security services, with a key focus on establishing and verifying secure communication. The methodology involves network-level security, where user requests are authenticated using a machine learning algorithm at the Defense Policy Unit (DPU). The approach includes data collection from the CSE-CIC-IDS2018 dataset, pre-processing steps such as dropping unnecessary columns and data standardization, and a combined weighted feature extraction and classification method. The system's effectiveness is evaluated using various performance parameters including accuracy, precision, recall, and F1-score, with results demonstrating the system's high effectiveness in detecting various types of network activities and attacks.

Keywords*:* Big Data, Cyber Security, Defense Policy Unit (DPU), Intrusion Detection System, Machine Learning.

1. Introduction

Big data refers to a new generation of technologies and architectures aimed at deriving value from large volumes of diverse data, characterized by high-velocity capture, discovery, and analysis. Originally defined by three key properties - volume, velocity, and variety - later expansions included additional characteristics: veracity, validity, value, variability, venue, vocabulary, and vagueness. Big data encompasses diverse formats like text, audio, images, and video, highlighting the importance of variety. Privacy in big data is maintained through different mechanisms at various stages of its life cycle: data generation, storage, and processing. During data generation, privacy is ensured through access restrictions and data falsification. In the storage phase, encryption methods (such as Identity Based Encryption, Attribute Based Encryption, and storage path encryption) and hybrid cloud setups are used for safeguarding sensitive information. The data processing stage employs Privacy Preserving Data Publishing (PPDP), which includes anonymization techniques like generalization and suppression, as well as methods like clustering, classification, and association rule mining for analyzing data. To manage big data's extensive volume, velocity, and variety, efficient frameworks are necessary for processing the vast amount of rapidly arriving data from diverse sources. Big data undergoes multiple phases in its life cycle, each requiring specific approaches to handle the data effectively. Therefore, privacy and security are major concern in big data. Big data security models are often not recommended for complex applications and can be disabled by default, leading to potential data compromises. Privacy involves the right to control how personal information is collected and used, aiming to protect individuals from unauthorized information disclosure. Security, on the other hand, focuses on defending data from unauthorized access, disclosure, disruption, and other threats through technology and processes. At the network level, big data security faces challenges due to the enormous volume and velocity of data generated, which complicates effective monitoring and analysis. The heterogeneity of big data, existing in various formats like logs, packets, and flow records, requires diverse tools and techniques for analysis and security. Additionally, ensuring data privacy, especially for personally identifiable information, is crucial to prevent unauthorized access in network-level big data. Therefore, this paper presents the Cyber Security Big Data

(CSBD) system is a cutting-edge approach for enhancing security in big data contexts. The key feature is its use of machine learning at the Defense Policy Unit (DPU) for authenticating user access requests. This involves identifying user data packets at the network level to ensure secure and authenticated user access to the CSBD system.

2. Literature Review

Rawat et al. [1] explored recent cyber security research related to big data, discussing how big data is protected and used for cyber security. Hababeh et al. [2] proposed a methodology to classify and secure big data before data mobility, duplication, and analysis. They focus on securing data mobility by classifying data into confidential and public categories, especially within the Hadoop Distributed File System. Gao et al. [3] introduced a Big Data Provenance Model (BDPM) for data supervision. This model supports various data types and processing modes, representing the data transformation process in big data systems and enriching provenance analysis functions. Awaysheh et al. [4] presented a security-by-design framework for Big Data frameworks deployment over cloud computing, named Big Cloud. It includes a systematic security analysis and automated security assessment, validated through an Apache Hadoop stack use case. Lin et al. [5] conducted a comprehensive overview of medical big data research, focusing on chronic diseases and health monitoring. They examine the full cycles of big data processing, including pre-processing, tools, algorithms, visualization, and security, and combine these technologies with specific medical needs. Sandhu et al. [6] discussed comparative analysis of cloud-based big data frameworks, addressing challenges in distributed database storage, security, heterogeneity, and visualization. Su et al. [7] presented a joint match-coalitional game-based security-aware resource allocation scheme for mobile social big data, showing superior performance in simulation experiments compared to existing schemes. Wang et al. [8] presented a framework for Big Data-as-a-Service, encompassing sensing, cloud, and application planes, and demonstrate its effectiveness using a tensor-based multiple clustering example on bicycle renting and returning data. Yang et al. [9] proposed a scalable data compression approach that calculates similarity among partitioned data chunks, using Map Reduce for algorithm implementation to enhance scalability on the Cloud. Babar et al. [10] suggested optimized IoT-enabled big data analytics architecture for edge-cloud computing with machine learning. This two-layered scheme (IoT–edge and cloud processing) uses optimized Map Reduce and YARN for efficient data processing and management, showing superior efficiency compared to existing methods. Alhazmi et al. [11] presented a big data security solution using block chain technology, incorporating fragmentation, encryption, and access control. Their fragmentation algorithm considers the data owner's encryption demands, offering significant security and privacy with minimal computational overheads. Dener et al. [12] proposed STLGBM-DDS, a DoS Intrusion Detection System for WSNs on Apache Spark. This system achieves high accuracy rates across various attack classes, indicating its effectiveness in detecting DoS attacks. Gao et al. [13] developed a provenance generation framework for log analysis, showing that their method can process files above MB level with minimal time overhead and generate accurate provenance information in near real-time. Long et al. [14] introduced a random attention-based data fusion approach for intrusion detection, demonstrating efficient model training and reduced false alarm rates on the CIC-IDS2018 dataset. He et al. [15] suggested an optimal hybrid relay selection scheme for enhanced security, using a weighting factor to ensure its effectiveness in various modes. Simulation results indicate improved security performance. Xu et al. [16] developed an improved CNN model for security performance prediction, combining different convolutional layers and inception blocks. This model shows a 20% increase in prediction precision over other methods. Zheng et al. [17] proposed DRSA-Net, a Dual-Route Structure-Adaptive Graph Network, to model the nonlinearity in tabular feature vectors without prior assumptions. It learns a sparse graph structure between variables and uses dual-route message passing to characterize interactions. Mhawi et al. [18] proposed a novel Ensemble Learning (EL) algorithm-based Network IDS model, utilizing a hybrid of Correlation Feature Selection and Forest Panellized Attributes (CFS–FPA) for efficient feature selection. The model employs AdaBoosting and bagging ensemble learning algorithms to enhance four classifiers: Support Vector Machine, Random Forest, Naïve Bayes, and K-Nearest Neighbour. These classifiers are applied using a voting average technique for aggregation. The model, evaluated in both binary and multi-class classification forms on the CICIDS2017 dataset, achieved an accuracy of 99.7%, a false-negative rate of 0.053, and a false alarm rate of 0.004, making it highly effective for IT-based organizations. Ruizhe et al. [19] proposed a hybrid intrusion detection system combining a CFS-DE feature selection algorithm and a weighted Stacking classification algorithm. The CFS-DE algorithm limits the dimension of features by finding the optimal subset, and the weighted Stacking algorithm adjusts weights of base classifiers based on their training performance. Tested on the NSL-KDD and CSE-CIC-IDS2018 datasets, the model showed high accuracy (87.44% on KDDTest+ and 99.87% on CSE-CIC-IDS2018), precision (89.09% on KDDTest+ and 99.88% on CSE-CIC-IDS2018), recall, and F1-scores. Mohammadi et al. [20] proposed an IDS using feature selection and clustering algorithms, specifically the feature grouping based on linear correlation coefficient (FGLCC) algorithm and cuttlefish algorithm (CFA). A decision tree is used as the classifier. Applied on the KDD Cup 99 dataset, their method demonstrated high accuracy (95.03%) and detection rate (95.23%), with a low false positive rate (1.65%).

3. Research Methodolgy

The Cyber Security Big Data (CSBD) system is a novel approach designed to enhance security in big data environments. Its primary function is to provide a foundational platform for selecting appropriate security services. One of its key features is the establishment and verification of secure communication within cyber security big data systems, ensuring robust protection. The architectural diagram, as shown in Figure 1, illustrates the structure and operational mechanism of the CSBD. In this methodology network level security is presented. Whenever any user wants to access the CSBD system then he/she would be authenticated. For authentication user data packets are identified at DPU using machine learning algorithm. Défense Policy Unit (DPU) is activated in this level that implements the machine learning approach for authenticating user request packets.

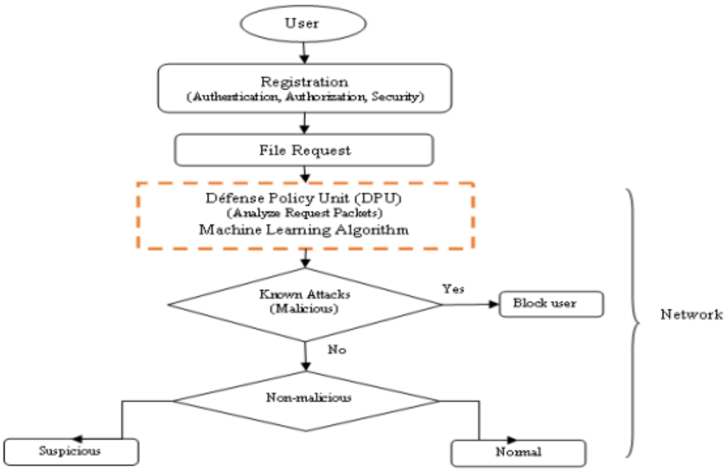
User
Registration
(Authentication, Authorization, Security)
File Request
Défense Policy Unit (DPU)
(Analyze Request Packets)
Machine Learning Algorithm
Known Attacks
(Malicious)
Yes
Block user
No
Non-malicious
Suspicious
Normal
Network

Fig. 1. Proposed Methodology

Working of Défense Policy Unit (DPU) is presented below in following steps:

Data Collection: In this step data is collected from CSE-CIC-IDS2018 dataset [21] [22].

Pre-processing: The pre-processing steps are used to clean the dataset for further modelling the machine learning model. The steps used in pre-processing are:

- Dropping Unnecessary Columns: In this step, some information in given dataset are unnecessary such as “Flow ID”, “SRC Port”. Therefore, these features are removed out from dataset.
- Function for Data Type Conversion: Then in this step, data conversion is performed. In this step, non-numeric entries are converted into appropriate data type such as into or float. This step is necessary as the machine learning model effectively process the numerical data as compared to alphanumeric data.
- Data Standardization: In this step, normalization is performed over converted data.

Combined Weighted Feature Extraction and Classification (CWFEC): In this step, the model presented a combined weighted feature extraction approach for CSBD security enhancement. It is evaluated as:

$$\mathrm{CWFEC} = \mathrm{w}_1 * \mathrm{F}_1 + \mathrm{w}_2 * \mathrm{F}_2 + \mathrm{w}_3 * \mathrm{F}_3 \qquad (1)$$

Where,w_1, w_2 and w_3 are weight factors and F_1 represents the feature set generated by PCA. F_2 represents the feature set by Pearson correlation and F_3 are the feature set generated out from random forest feature importance. The common features extracted out are combined together and fed into random forest classifier for classification of attacks.

4. Data Analysis

Dataset Description: In this paper, CSE-CIC-IDS2018 dataset is used [21][22]. The CSE-CIC-IDS2018 dataset is a joint effort by the Communications Security Establishment (CSE) and the Canadian Institute for Cyber security (CIC) created for intrusion detection purposes [21]. This dataset is an enhancement over the previous CIC-IDS2017 version, featuring a larger scale with 16,233,002 records, each containing 80 features. These records were generated through simulations involving attacks from 50 computers. However, it's noted that the dataset contains numerous redundant features which might not be highly relevant for intrusion detection, potentially impacting the performance of such systems.

Performance Parameters: In this paper, following performance parameters are used:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad (2)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (3)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (4)$$

$$\mathrm{F1\_Score} = \frac{2*\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision+Recall}} \qquad (5)$$

The fig 2 presents the accuracy of different models in detecting various types of network activities or attacks. The "Normal" model, which identifies standard network behaviour, shows a high accuracy of 95%. The "Bot" model, tailored to detect bot-related activities, has a slightly lower accuracy of 93%. The model designed to identify Denial of Service (DoS) attacks has an accuracy of 91%. Lastly, a model for detecting other attacks with an accuracy of 92%. The fig 3 summarizes the precision of different models. The model for normal activities has a precision of 95%, the bot detection model has 93%, the DoS attack model has 91%, and the model for other unspecified attacks has a precision of 94%. The fig 4 provides recall metrics for models designed to detect different types of network activities or attacks. The model for normal activities has a recall of 94%, indicating it accurately identifies 94% of normal activities. The bot detection model has a recall of 92%, the DoS attack model 93%, and the model for other unspecified attacks also 93%. These figures suggest that each model is highly effective in correctly identifying the specific type of activity or attack it's designed to detect, with a particular strength in accurately recognizing true instances of each category. The fig 5 summarizes the F1-scores of various models. The F1-score, a harmonized measure of precision and recall, indicates the models' accuracy and reliability. The model for normal network activities achieves a high F1-score of 95, reflecting its excellent accuracy in identifying standard behaviours. The bot detection model has a slightly lower but still strong score of 93, demonstrating its effectiveness in identifying bot-related activities. The model for detecting Denial of Service (DoS) attacks shows an F1-score of 92, indicating a high level of accuracy in this specific area. Lastly, the model designed to identify other unspecified types of attacks has an F1-score of 94, suggesting it is very adept at accurately detecting a broad range of potential threats.

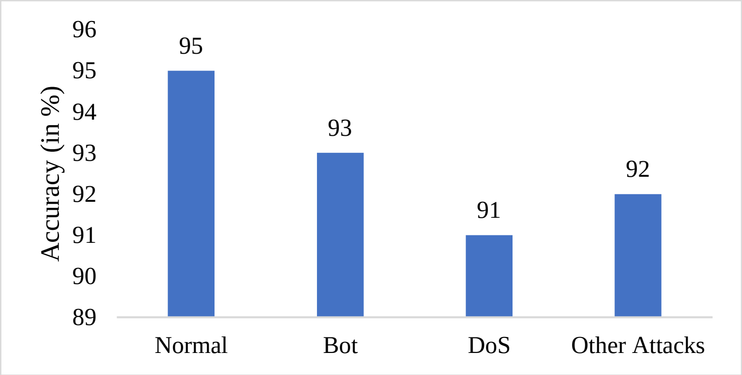

Figure 2. Accuracy Evaluation

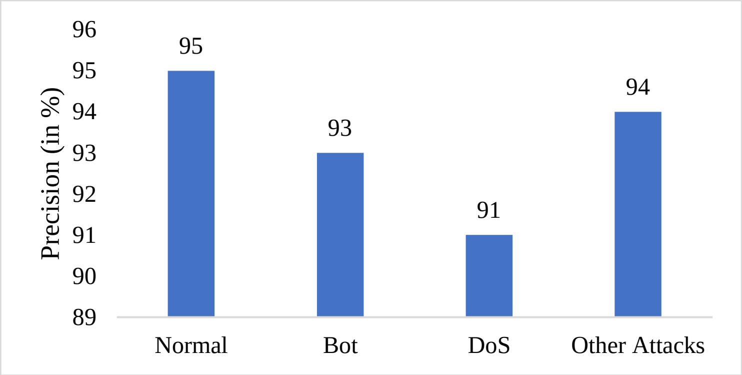

Figure 3. Precision Evaluation

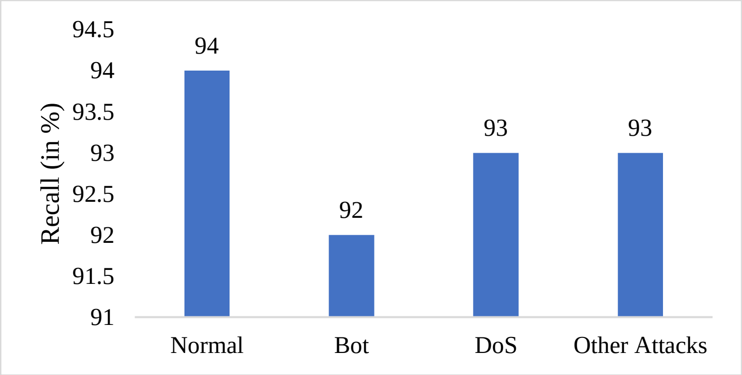

Figure 4. Recall Evaluation

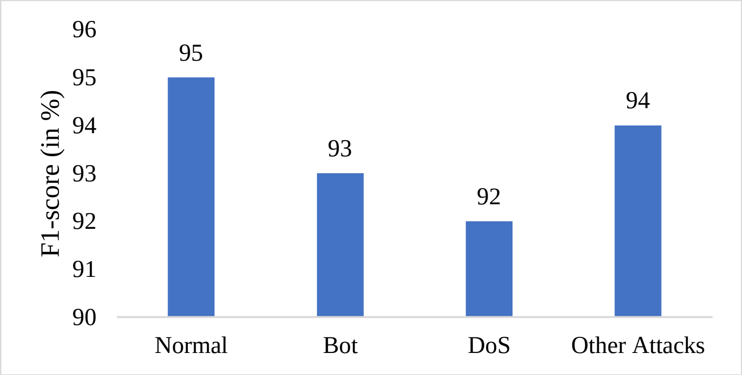

Figure 5. F1-Score Evaluation

5. Conclusion

The paper demonstrates the evolving landscape of intrusion detection systems for network security in big data environment. The literature review explored the feature extraction techniques and their capabilities for advanced machine learning techniques, such as Ensemble Learning, weighted stacking algorithms, etc. These approaches presented a significant effectiveness of IDS in cyber security. Motivated by these approaches, the paper presented a combined weighted feature extraction approach for network layer security in big data. The CSBD system, through its innovative use of machine learning algorithms and data pre-processing techniques, has demonstrated significant efficiency in enhancing cyber security in big data environments. The performance evaluations, as illustrated in the various figures, indicate high accuracy, precision, recall, and F1-scores across different models designed for detecting normal activities, bots, DoS attacks, and other attacks. These results highlight the system's robustness and reliability in accurately identifying and distinguishing between normal and malicious activities. The implementation of the CSBD system offers a promising direction for future advancements in cyber security, particularly in handling the complexities associated with big data.

References

[1] D. B. Rawat, R. Doku, and M. Garuba, “Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security,” IEEE Trans. Serv. Comput., vol. 14, no. 6, pp. 2055–2072, 2021, doi: 10.1109/TSC.2019.2907247.

[2] I. Hababeh, A. Gharaibeh, S. Nofal, and I. Khalil, “An Integrated Methodology for Big Data Classification and Security for Improving Cloud Systems Data Mobility,” IEEE Access, vol. 7, pp. 9153–9163, 2019, doi: 10.1109/ACCESS.2018.2890099.

[3] Y. Gao, X. Chen, and X. Du, “A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model,” IEEE Access, vol. 8, pp. 38742–38752, 2020, doi: 10.1109/ACCESS.2020.2975820.

[4] F. M. Awaysheh, M. N. Aladwan, M. Alazab, S. Alawadi, J. C. Cabaleiro, and T. F. Pena, “Security by Design for Big Data Frameworks Over Cloud Computing,” IEEE Trans. Eng. Manag., vol. 69, no. 6, pp. 3676–3693, 2022, doi: 10.1109/TEM.2020.3045661.

[5] R. Lin, Z. Ye, H. Wang, and B. Wu, “Chronic Diseases and Health Monitoring Big Data: A Survey,” IEEE Rev. Biomed. Eng., vol. 11, pp. 275–288, 2018, doi: 10.1109/RBME.2018.2829704.

[6] A. K. Sandhu, “Big data with cloud computing: Discussions and challenges,” Big Data Min. Anal., vol. 5, no. 1, pp. 32–40, 2022, doi: 10.26599/BDMA.2021.9020016.

[7] Z. Su and Q. Xu, “Security-Aware Resource Allocation for Mobile Social Big Data: A Matching-Coalitional Game Solution,” IEEE Trans. Big Data, vol. 7, no. 4, pp. 632–642, 2021, doi: 10.1109/TBDATA.2017.2700318.

[8] X. Wang, L. T. Yang, H. Liu, and M. J. Deen, “A Big Data-as-a-Service Framework: State-of-the-Art and Perspectives,” IEEE Trans. Big Data, vol. 4, no. 3, pp. 325–340, 2018, doi: 10.1109/TBDATA.2017.2757942.

[9] C. Yang and J. Chen, “A Scalable Data Chunk Similarity Based Compression Approach for Efficient Big Sensing Data Processing on Cloud,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 6, pp. 1144–1157, 2017, doi: 10.1109/TKDE.2016.2531684.

[10] M. Babar, M. A. Jan, X. He, M. U. Tariq, S. Mastorakis, and R. Alturki, “An Optimized IoT-Enabled Big Data Analytics Architecture for Edge–Cloud Computing,” IEEE Internet Things J., vol. 10, no. 5, pp. 3995–4005, 2023, doi: 10.1109/JIOT.2022.3157552.

[11] H. E. Alhazmi, F. E. Eassa, and S. M. Sandokji, “Towards Big Data Security Framework by Leveraging Fragmentation and Block chain Technology,” IEEE Access, vol. 10, pp. 10768–10782, 2022, doi: 10.1109/ACCESS.2022.3144632.

[12] M. Dener, S. Al, and A. Orman, “STLGBM-DDS: An Efficient Data Balanced DoS Detection System for Wireless Sensor Networks on Big Data Environment,” IEEE Access, vol. 10, pp. 92931–92945, 2022, doi: 10.1109/ACCESS.2022.3202807.

[13] Y. Gao, X. Chen, B. Li, and X. Du, “A Near Real-Time Big Data Provenance Generation Method Based on the Conjoint Analysis of Heterogeneous Logs,” IEEE Access, vol. 11, pp. 80806–80821, 2023, doi: 10.1109/ACCESS.2023.3300844.

[14] J. Long, W. Liang, K.-C. Li, Y. Wei, and M. D. Marino, “A Regularized Cross-Layer Ladder Network for Intrusion Detection in Industrial Internet of Things,” IEEE Trans. Ind. Informatics, vol. 19, no. 2, pp. 1747–1755, 2023, doi: 10.1109/TII.2022.3204034.

[15] H. He, P. Ren, Q. Du, L. Sun and Y. Wang, "Enhancing Physical-Layer Security via Big-Data-Aided Hybrid Relay Selection," in Journal of Communications and Information Networks, vol. 2, no. 1, pp. 97-110, March 2017, doi: 10.1007/s41650-017-0008-8.

[16] L. Xu, X. Zhou, Y. Tao, L. Liu, X. Yu, and N. Kumar, “Intelligent Security Performance Prediction for IoT-Enabled Healthcare Networks Using an Improved CNN,” IEEE Trans. Ind. Informatics, vol. 18, no. 3, pp. 2063–2074, 2022, doi: 10.1109/TII.2021.3082907.

[17] Q. Zheng et al., “Deep Tabular Data Modelling with Dual-Route Structure-Adaptive Graph Networks,” IEEE Trans. Know. Data Eng., vol. 35, no. 9, pp. 9715–9727, 2023, doi: 10.1109/TKDE.2023.3249186.

[18] Mhawi, Doaa N., Ammar Aldallal, and Soukeana Hassan. "Advanced feature-selection-based hybrid ensemble learning algorithms for network intrusion detection systems." Symmetry 14.7 (2022): 1461.

[19] Zhao, Ruizhe, et al. "A hybrid intrusion detection system based on feature selection and weighted stacking classifier." IEEE Access 10 (2022): 71414-71426.

[20] Mohammadi, Sara, et al. "Cyber intrusion detection by combined feature selection algorithm." Journal of information security and applications 44 (2019): 80-88.

[21] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1 (2018): 108-116.

[22] https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv/data