Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Leveraging of Gradient Boosting Algorithm in Misuse Intrusion Detection using KDD Cup 99 Dataset

Sulaiman Muhammed Sulaiman[1*], Adnan Mohsin Abdulazeez[1]
sulaiman.muhammed@dpu.edu.krd , adnan.mohsin@dpu.edu.krd
[1]ITM Dept., Technical Institute of Administration-Duhok, Duhok Polytechnic University, Duhok, Iraq

Article Information

Submitted : 23 Jan 2024
Reviewed: 26 Jan 2024
Accepted : 10 Feb 2024

Keywords

Intrusion Detection, Ensemble Learning, Adaboost, Lightgbm, and XGBoost

Abstract

This study addresses the persistent challenge of intrusion detection as a long-term cybersecurity issue. Investigating the efficacy of machine learning algorithms in anomaly and misuse detection. Research employs supervised learning for misuse detection and explain anomaly detection. Focused on adaptability and continual evolution the study explores the application of ensemble learning models AdaBoost, LightGBM, and XGBoost. Applying these algorithms in the context of intrusion detection. Utilizing the KDD Cup 99 dataset as a benchmark the paper assesses and compares the performance of these models. Besides, illuminating their effectiveness particularly in identifying smurf attacks within the cybersecurity landscape.

A. Introduction

Given of the rapid technology improvements cybersecurity emerges as a pivotal discipline grappling with the complexities of an evolving digital landscape. The surge in interconnected systems and the escalating sophistication of cyber threats underscore the urgent need to fortify networks against threat of intrusions. Intrusion detection a linchpin of cybersecurity assumes a central role in identifying and thwarting unauthorized access. This imperative function contributes to preserving the integrity and security of digital assets safeguarding against potential compromise and threat [1].

The KDD Cup 99 dataset is a seminal benchmark in the field of intrusion detection, widely recognized for evaluating the efficacy of cybersecurity systems. Compiled from a variety of network activities it serves as a comprehensive simulation of real-world cyber threats. The dataset encompasses diverse attack scenarios offering a rich and representative environment for assessing intrusion detection mechanisms. Introduced during the Third International Knowledge Discovery and Data Mining Tools Competition. This dataset has been instrumental in fostering research and innovation in the development and evaluation of intrusion detection systems, making it a cornerstone in academic investigations within the cybersecurity domain [3,4].

The integration of machine learning algorithms in intrusion detection represents a significant advancement in the space of cybersecurity research. Machine learning techniques including supervised and unsupervised learning are applied to discern patterns of normal and malicious activities within network data. Supervised learning models trained on labeled datasets exhibit proficiency in recognizing known intrusion patterns, while unsupervised learning models excel in identifying anomalous behavior indicative of novel threats. The utilization of machine learning contributes to the development of adaptive intrusion detection systems capable of continuously learning and evolving to confront the dynamic landscape of cyber threats. This academic endeavor explores the nuanced application of machine learning algorithms emphasizing their role in enhancing the accuracy and efficiency of intrusion detection methodologies [5].

In an era where cyber threats are dynamic and ever-evolving, understanding the capabilities of machine learning models in detecting network intrusions becomes paramount [6]. This study endeavors to contribute to the expanding library of research in cybersecurity by providing insights into the performance of ensemble learning models. Thus, empowering cybersecurity practitioners and researchers in their ongoing efforts to fortify digital ecosystems against malicious incursion.

Our study assesses the effectiveness of ensemble learning models AdaBoost, LightGBM, and XGBoost in the intricate space of intrusion detection, examining their performance in discerning nuanced cyber threats.

B. Literature View

In [7], Gad, A. R and etl. designed a monitor system applications and network traffic to detect suspicious activities and raise alerts when potential threats are identified. the researcher focuses on applying machine learning

algorithms such as Support Vector Machine (SVM) and Artificial Neural Networks (ANN), to detect intrusion rates. In order to improve the accuracy of intrusion detection. The authors utilized feature selection techniques and Chi-Squared Based algorithms, with NSL KDD dataset. The experimentation with dataset resulted in approximately 48% accuracy for the SVM algorithm and 97% accuracy for the ANN algorithm, highlighting the superior performance of the ANN model in this context.

A. Mohamed and other in [8], discusses the application of machine learning techniques in enhancing online security through intrusion detection systems (IDS). UNSW-NB15 represents as dataset to analyze and compare various machine learning strategies, such as decision trees, support vector machines (SVM), random forests, and deep learning models. The dataset includes benign and malicious network activity examples, and after preprocessing and feature engineering, the study evaluates the constructed models using metrics like accuracy, precision, recall, and F1 score. The findings demonstrate that different machine learning algorithms can effectively detect various types of attacks in the dataset, and the research also examines the impact of different feature engineering strategies on detection accuracy. Results of this study the showed the efficiency of using machine learning algorithms in designing intrusion detection systems.

In [9] Ugochukwu, C. J. present suggestion for implementing machine learning in intrusion detection. The main findings of the paper are using Random Forest and Random Tree algorithms were the most efficient in classifying the attacks on the Test dataset, with precision and F-measure above 97%. Bayes Net outperformed other algorithms in terms of detection rate. The paper recommends further research to explore other machine learning algorithms for improved classification efficiency [9].

Maseer, Z. K and etl study, demonstrates the successful use of machine learning algorithms for multi-class classification tasks in network intrusion detection systems of attacks types (DDoS, PROBE, R2L, and U2R), achieving an accuracy of 95.95%. The proposed method focuses on detecting anomalies and protecting the SDN platform from attacks in real-time scenarios. The algorithm used in this study are Decision Tree, Random Forest and XGBoost. The dataset analyzed for this system is NSL-KDD dataset [10].

Exploring the advanced threat attacks, and the drawbacks of traditional network intrusion detection systems based on feature filtering because it has limitations in effectively thwarting of these attacks are explained by P. V. Pandit in [11]. They addressed machine learning techniques such as neural networks, statistical models rule learning and ensemble methods, are being used to create more effective intrusion detection systems. The suggest a novel ensemble method, combining decision tree, random forest, extra tree, and XGBoost algorithms, was proposed in this study to improve intrusion detection accuracy. The method was implemented in Python and evaluated using the CICIDS2017 dataset, showing a significant increase in detection accuracy.

In [12], the paper focused on of an intrusion detection system using AIS-ELM, the goal of applying this approach to a smart home network gateway, and the proposal of AIS-ELM computational techniques for intrusion detection systems in smart homes. ELM, AIS, Clonal Algorithm used for proposed model. The dataset

used in the study is the network traffic generated from the Mozilla Gateway controlled smart home system. This dataset is used to train the AIS-ELM based IDS to detect anomalies in a smart home network.

The main findings of the paper [13], are presenting a comprehensive survey on intrusion prevention and detection using neural networks for data security. In result discovered that anomaly detection is the best method for Intrusion detection based on feature selection. Also, need for further research on intrusion detection systems using machine learning and neural networks. Dataset used in this proposal for network intrusion are containing 22 out of 29 types of attacks, divided into training and validation sets, and used for K-means clustering and SVM with Deep learning for intrusion identification. It also includes key features such as number of data values, epoch, loss, accuracy, runtime, number of false positive, and number of false negative.

C. Smurf Attack Hazard

The distinctive characteristic of smurf attacks lies in their utilization of the Internet Control Message Protocol (ICMP) to inundate a target network with echo request (ping) packets distinguishing them from other forms of network intrusions. Originating from a malicious virus these attacks exploit the broadcast nature of ICMP queries [14,15]. Then it amplifying their impact by disseminating requests to broadcast addresses across multiple networks. The ensuing surge of responses engenders network congestion, service degradation, and occasional outages for the targeted system. Despite their nomenclature drawing from 1990s cartoon characters, smurf attacks persist as a cybersecurity threat evolving over time and continuing to pose challenges to network security [16]. Understanding the fundamental processes of such attacks is crucial for mitigating them and strong intrusion detection systems that can identify the minute indicators of smurf attacks amidst the vast sea of network traffic. In our study smurf attacks serve as the focal point for evaluating the effectiveness of ensemble learning models offering valuable insights into the models' capacity to identify and counteract this specific class of cyber threat.

Moreover, the hazard posed by Smurf attacks is exacerbated by their evolution over time and persistence as a cybersecurity threat. Smurf attacks have adapted to contemporary network environments, necessitating continuous vigilance and advanced intrusion detection mechanisms. The dynamic nature of these attacks underscores the ongoing challenge in mitigating their impact and reinforces the critical importance of robust cybersecurity strategies to safeguard against such hazardous intrusions. The abstraction introduced in this study allows for a controlled examination of key concepts related to Smurf attacks, facilitating clarity in the presentation of analytical methods and findings. However, the complexity inherent in genuine network traffic data remains essential for extrapolating these concepts to real-world scenarios and refining intrusion detection strategies in the face of evolving cybersecurity threats.

To get further insight, we conduct a smurf assault experiment and observe the potential hazards via ICMP ping. In figure below the term "time intervals" pertains to distinct temporal as a segments, while "packet counts" denotes the numerical representation of packets received within those intervals. It is

important to note that the data utilized in this context does not constitute authentic network traffic information rather it serves as a simplified representation for illustrative purposes. In reality, a comprehensive analysis and visualization of patterns associated with a Smurf assault demand authentic network traffic data. Such actual network data, reflecting the intricacies of real-world network activities, is indispensable for a nuanced understanding of Smurf attacks and the development of effective intrusion detection methodologies.

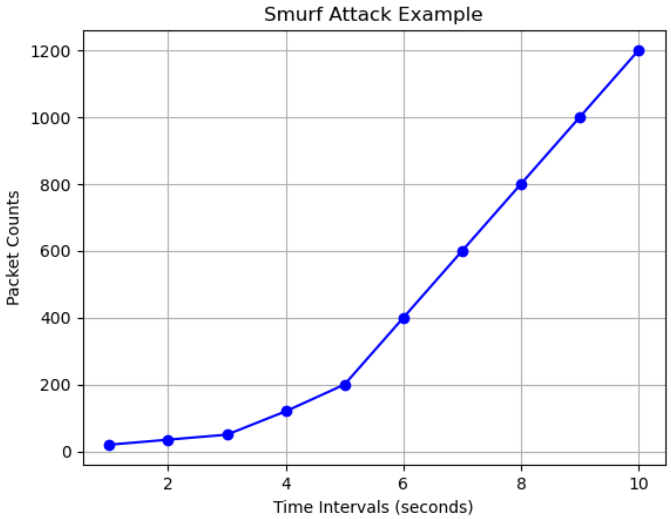

Figure 1. Smurf Attack Example [1]

An effective class of machine learning methods called gradient boosting algorithms is intended for regression and predictive modeling applications. These algorithms are fundamentally based on the capacity to progressively construct an ensemble of weak learners, which are usually decision trees, and then aggregate their predictions to form a reliable and accurate model. The primary innovation that sets Gradient Boosting apart from conventional techniques is its emphasis on error reduction through the progressive fitting of new models to the residuals of the previous models. Gradient Boosting models thrive at collecting intricate associations in the data because of this iterative process, which makes them very useful in situations where high predicted accuracy is very essential [17,18].

The minimization of a loss function, which gauges the discrepancy between expected and actual values, is the basic idea underlying gradient boosting. By iteratively changing the model parameters in the direction that minimizes the loss, the method optimizes the parameters.

Gradient Boosting Machines (GBM), XGBoost, and LightGBM are popular gradient boosting implementations that each provide special improvements and functionality. For instance, XGBoost uses parallel processing and regularization terms to increase efficiency, whereas LightGBM optimizes the construction of decision trees using histogram-based learning to increase speed and scalability [19]. Collectively, Gradient Boosting algorithms have become indispensable tools in machine learning, frequently used across various domains such as finance, healthcare, and cybersecurity due to their exceptional predictive performance and adaptability [20].

D. Identifying Introversion Utilizing ML

The primary objective of intrusion detection by machine learning (IDML) is a crucial component from assignments of cybersecurity is to locate and address harmful activity occurring within all of computer system or network. Conventional intrusion detection techniques in previous studies as on frequently depend predominantly on the signature on another hand on rule-based systems, which in turn can be for most cases difficult to adjust to complex also ever-evolving assault patterns.

In contrast, machine learning makes use of sophisticated statistical and computational methods to automatically identify and learn from patterns that point to malevolent activity. This paradigm change provides a method to intrusion detection that is dynamic and adaptable, able to handle the subtleties and complexity of contemporary cyberthreats. Anomaly detection and abuse detection are the two main types of machine learning models used in intrusion detection. [21,22]

1) Anomaly detection involves building a model of normal behavior and flagging any deviation from this norm as a potential intrusion.
2) Misuse detection, on the other hand, relies on predefined patterns of known attacks and malicious activities. Supervised learning, where models are trained on labeled datasets with examples of both normal and malicious behavior, is commonly employed for misuse detection. Unsupervised learning, including clustering and outlier detection, is often utilized for anomaly detection, where the model learns to identify patterns without explicit labeling.

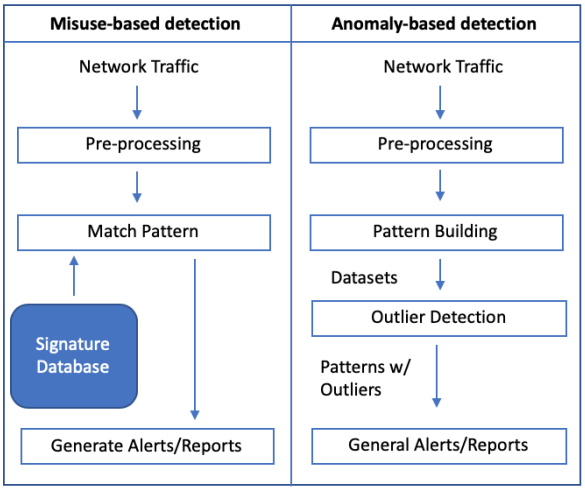

Figure 2. Misuse Vs Anomaly Detection [2]

The effectiveness of intrusion detection by machine learning lies in its ability to continuously adapt and improve its detection capabilities. As cyber threats evolve and become more sophisticated, machine learning models can be retrained with new data to stay current [24,25]. Additionally, feature engineering plays a crucial role, involving the selection and transformation of relevant input features to enhance the model's ability to discriminate between normal and malicious

activities. The application of ensemble techniques, such as combining multiple models to make collective decisions, further enhances the robustness and accuracy of intrusion detection systems [27]. Despite the advancements in machine learning-based intrusion detection, challenges persist. Adversarial attacks, where malicious actors intentionally manipulate data to deceive the model, pose a significant threat. Ensuring the privacy and security of the training data is another concern, especially when dealing with sensitive information [28,29].
Ongoing research and development in machine learning techniques, coupled with a comprehensive understanding of cybersecurity threats, are essential for creating effective and resilient intrusion detection systems in the ever-evolving landscape of cyber threats [30].

E. Methodology:

The research methodology involves preprocessing the KDD Cup 99 dataset encoding categorical features and splitting the data into training and testing sets. Three ensemble learning models AdaBoost, LightGBM, and XGBoost are chosen for their versatility and success across various domains. The training process and subsequent evaluation using accuracy as a metric form the core of the methodology.

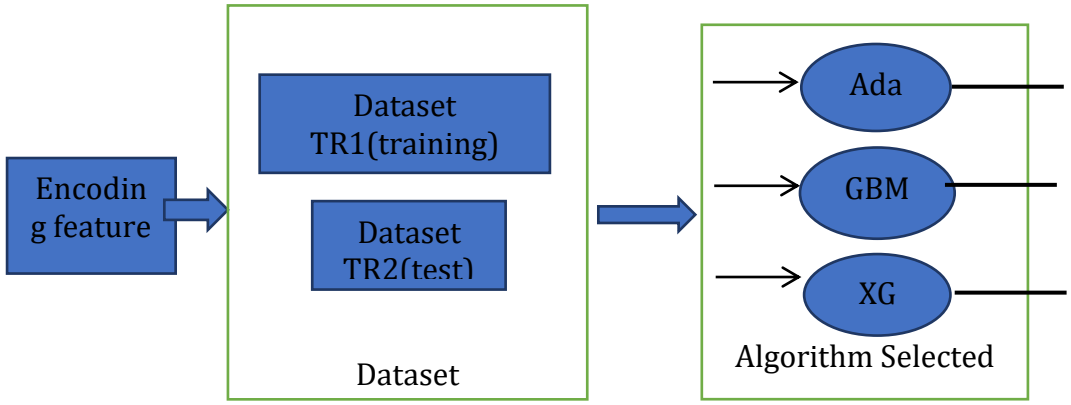

Figure 3. Implementing ML on KDD Cup 99 dataset [3]

A) AdaBoost

AdaBoost worked by Sequential training corrects errors by assigning higher weights to misclassifications. This adaptive weighting of misclassified instances allows AdaBoost to iteratively refine its model, emphasizing challenging data points. The final model is an ensemble of these weak learners, collectively forming a robust and accurate predictor. In the context of the code, AdaBoost contributes to the ensemble of models for detecting smurf attacks within the dataset. Top of Form
Bottom of Form

B) LightGBM

In another hand, LightGBM optimizes decision tree construction using histogram-based learning, prioritizing nodes with larger data contributions. This approach accelerates training and enhances scalability, making it effective for the KDD Cup 99 dataset.

C) XGBoost

employs a incorporates regularization terms to control model complexity and parallel processing to enhance efficiency. The model is trained iteratively, adjusting weights and combining weak learners to create a robust and accurate predictive model. In the context of intrusion detection, XGBoost contributes to the ensemble of models aimed at identifying smurf attacks within the dataset.

F. Results and Discussion:

This section shows experiments compelling results in term of algorithm selected and result obtained from this experiment. AdaBoost exhibits [0.99076], emphasizing its adaptability to diverse datasets. LightGBM, designed for efficient handling of large datasets, achieves [0.99925] highlighting its competitive performance. XGBoost, a widely employed algorithm, demonstrates robustness with an accuracy of [0.99985].

Table 1. Comparison of Gradient Boosting

	Accuracy	Metrics
1	XGBoost	0.999850976
2	GBM	0.999254881
3	AdaBoost	0.990760519

The nuanced differences in accuracy underscore the importance of model selection in intrusion detection. dive into the observed performance differences. AdaBoost's ability to adapt to different datasets is evident, making it a versatile choice. LightGBM's efficiency in processing large datasets is a notable advantage, contributing to its competitive accuracy. XGBoost, although slightly trailing in accuracy, showcases robustness and generalization capabilities. The trade-offs between these models are discussed, providing insights for practitioners in choosing the most suitable model for their intrusion detection needs.

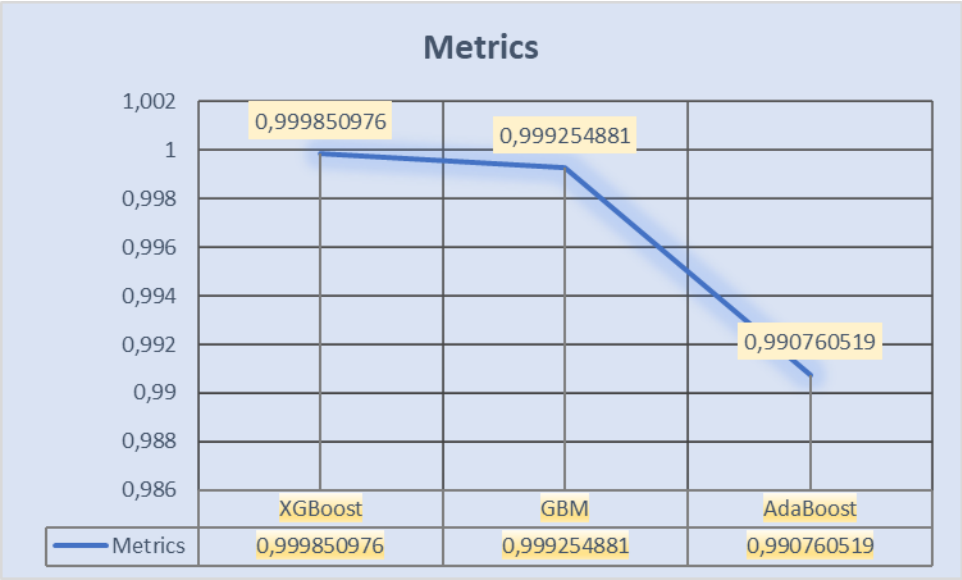

Figure 4. Accuracy of Algorithms selected [4]

G. Future Work

The implications of the study extend to the space of cybersecurity, emphasizing the significance of machine learning in intrusion detection. Future work may involve fine-tuning hyperparameters, exploring additional feature engineering strategies, and assessing the models' performance across diverse intrusion scenarios.

H. Conclusion

In conclusion, this study presents a comparative analysis of AdaBoost, LightGBM, and XGBoost for smurf attack detection using the KDD Cup 99 dataset. The nuanced differences in accuracy underscore the importance of tailored model selection in intrusion detection applications. This research contributes to the evolving landscape of cybersecurity and lays the groundwork for further exploration into Machine learning for intrusion detection. However, challenges such as adversarial attacks and ensuring data privacy remain, necessitating ongoing research and development efforts. The effectiveness of machine learning-based intrusion detection systems is contingent on robust feature engineering, thoughtful selection of algorithms, and vigilant model monitoring. Despite these challenges, the potential for significantly improving detection accuracy and response times is evident.

I. References

[1]. Kalimuthan, C., & Renjit, J. A. (2020). Review on intrusion detection using feature selection with machine learning techniques. *Materials Today: Proceedings*, *33*, 3794-3802.

[2]. Serinelli, B. M., Collen, A., & Nijdam, N. A. (2020). Training guidance with kdd cup 1999 and nsl-kdd data sets of anidinr: Anomaly-based network intrusion detection system. *Procedia Computer Science*, *175*, 560-565.

[3]. Meryem, A., & Ouahidi, B. E. (2020). Hybrid intrusion detection system using machine learning. *Network Security*, *2020*(5), 8-19.

[4]. Gümüşbaş, D., Yıldırım, T., Genovese, A., & Scotti, F. (2020). A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. *IEEE Systems Journal*, *15*(2), 1717-1731.

[5]. Konstantinov, A. V., & Utkin, L. V. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems*, *222*, 106993.

[6]. Saranya, T., Sridevi, S., Deisy, C., Chung, T. D., & Khan, M. A. (2020). Performance analysis of machine learning algorithms in intrusion detection system: A review. *Procedia Computer Science*, *171*, 1251-1260.

[7]. Gad, A. R., Nashat, A. A., & Barkat, T. M. (2021). Intrusion detection system using machine learning for vehicular ad hoc networks based on ToN-IoT dataset. *IEEE Access*, *9*, 142206-142217.

[8]. A. Mohamed, J. Heilala and N. S. Madonsela, "Machine Learning-Based Intrusion Detection Systems for Enhancing Cybersecurity," *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Singapore, Singapore, 2023, pp. 366-370,

[9]. Ugochukwu, C. J., Bennett, E. O., & Harcourt, P. (2019). *An intrusion detection system using machine learning algorithm*. LAP LAMBERT Academic Publishing.
[10]. Alzahrani, A. O., & Alenazi, M. J. (2021). Designing a network intrusion detection system based on machine learning for software defined networks. *Future Internet*, *13*(5), 111.

[10]. Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. *IEEE access*, *9*, 22351-22370.

[11]. P. V. Pandit, S. Bhushan and P. V. Waje, "Implementation of Intrusion Detection System Using Various Machine Learning Approaches with Ensemble learning," 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2023, pp. 468-472,

[12]. E. D. Alalade, "Intrusion Detection System in Smart Home Network Using Artificial Immune System and Extreme Learning Machine Hybrid Approach," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 2020, pp. 1-2.

[13]. V. Bhatia, S. Choudhary and K. R. Ramkumar, "A Comparative Study on Various Intrusion Detection Techniques Using Machine Learning and Neural Network," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2020.

[14]. Revathy, G., Rajendran, V., Sathish Kumar, P., Vinuharini, S., & Roopa, G. N. (2022, May). Smurf attack using hybrid machine learning technique. In *AIP Conference Proceedings* (Vol. 2463, No. 1). AIP Publishing.

[15]. Karpinski, M., Shmatko, A., Yevseiev, S., Jancarczyk, D., & Milevskyi, S. (2021). Detection Of Intrusion Attacks Using Neural Networks.

[16]. Shanker, R., Agrawal, P., Singh, A., & Bhatt, M. W. (2023). Framework for identifying network attacks through packet inspection using machine learning. *Nonlinear Engineering*, *12*(1), 20220297.

[17]. Singh, N. K., & BJ, S. K. (2023, July). Detection and Prevention of UDP Protocol Exploiting and Smurf Attack in WSN Using Sequential Probability Ratio Test

Algorithm. In *2023 International Conference on Data Science and Network Security (ICDSNS)* (pp. 1-6). IEEE.

[18]. N. K. Singh and S. K. B. J, "Detection and Prevention of UDP Protocol Exploiting and Smurf Attack in WSN Using Sequential Probability Ratio Test Algorithm," *2023 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 2023, pp. 1-6.

[19]. P. Ndayishimiyepas, C. Wilson and M. Kimwele, "A Hybrid Model for Predicting Missing Records in Data Using XGBoost," *2022 IEEE International Symposium on Product Compliance Engineering - Asia (ISPCE-ASIA)*, Guangzhou, Guangdong Province, China, 2022, pp. 1-5.

[20]. Salih, A. A., & Abdulazeez, A. M. (2021). Evaluation of classification algorithms for intrusion detection system: A review. *Journal of Soft Computing and Data Mining*, *2*(1), 31-40.

[21]. Alweshah, M., Hammouri, A., Alkhalaileh, S., & Alzubi, O. (2023). Intrusion detection for the internet of things (IoT) based on the emperor penguin colony optimization algorithm. *Journal of Ambient Intelligence and Humanized Computing*, *14*(5), 6349-6366.

[22]. Amanoul, S. V., Abdulazeez, A. M., Zeebare, D. Q., & Ahmed, F. Y. (2021, June). Intrusion detection systems based on machine learning algorithms. In *2021 IEEE international conference on automatic control & intelligent systems (I2CACIS)* (pp. 282-287). IEEE.

[23]. Al Tawil, A., & Sabri, K. E. (2021, July). A feature selection algorithm for intrusion detection system based on moth flame optimization. In *2021 International Conference on Information Technology (ICIT)* (pp. 377-381). IEEE.

[24]. Li, L., Zhang, S., Zhang, Y., Chang, L., & Gu, T. (2019, June). The intrusion detection model based on parallel multi-artificial bee colony and support vector machine. In *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)* (pp. 308-313). IEEE.

[25]. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), 1-22.

[26]. Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *Ieee Access*, *7*, 41525-41550.

[27]. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, *1*, 108-116.

[28]. Zhang, D., & Gong, Y. (2020). The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure. *IEEE Access*, *8*, 220990-221003.

[29]. Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24-39.

[30]. Leevy, J. L., Hancock, J., Zuech, R., & Khoshgoftaar, T. M. (2020, October). Detecting cybersecurity attacks using different network features with lightgbm and xgboost learners. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 190-197). IEEE.