Kasu Journal of Computer Science Vol. 1 No. 3 [September, 2024], pp. 489-496
https://doi.org//10.47514/kjcs/2024.1.3.006

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Improved Machine Learning Algorithm for Detecting Intrusions in Web-based Applications

YORO Rume Elizabeth[1] and EDUN Ogechi Peace[2]
[1, 2] Department of Cyber Security, Dennis Osadebay University Asaba, Nigeria
elizabeth.yoro@dou.edu.ng[1], edun.ogechi@dou.edu.ng[2]

Abstract

Background*: Intrusion is defined as an unauthorized access to sensitive information or data. When this happens, security is breached and confidentiality of the affected system is in doubt. With the proliferation of web applications, intrusion detection systems have increasingly become relevant in today's digital age.* ***Aim:*** *Existing intrusion detection systems have low accuracy and precision rates which causes false alarm. Therefore, the aim of this research is to develop an improved algorithm that has the capacity to correctly detect intruders at the right time* ***Method****: Real-world datasets collected from a forensic framework was used to evaluate the performance of the proposed algorithm. The dataset was preprocessed to remove redundant or irrelevant features which were normalized. The datasets were then divided into training and testing sets.* ***Results****: The proposed algorithm outperformed existing ones with accuracy of 98.89%. Research findings suggest that, the proposed algorithm is effective in detecting intrusions in web-based applications.*

Keywords: *intrusion, intrusion detection, datasets, algorithm, machine learning*

1. Introduction

Intrusion is an unauthorized access to data or systems deployed over a computer network. It is the process of ensuring safety, confidentiality and integrity of computing resources. With the proliferation of web-based applications, demand for intrusion detection systems is on high toll. These are systems with the capacity of detecting suspicious elements based on user's profiles. When these malicious activities are detected as soon as attempts occur, criminal intentions or security breaches are averted (Jayakumar et al., 2021).

Previously, Network Intrusion Detection Systems (NIDS) were utilized as shield for resources deployed over a network. Consequently, NIDS was used to detect breaches in network environment (Sirajuddin et al., 2021; Jayakumar et al., 2021). Many datasets were created to validate NIDS. However, these datasets were adjudged to be inadequate for use in real-world environment. This led to the development of robust network datasets by Researchers to reliably and correctly predicts intrusion and other forms of breaches (Yaacoub et al., 2020). Furthermore, Host Intrusion Detection Systems (HIDS) has capacity of detecting intrusion on one web host and devise at a time (Satilmiş et al., 2023). HIDS detects intrusion based on analysis of system calls, logs, linked library files and registry keys.

Organizations rely on intrusion detection system (IDS) to avoid invasions. This is a crucial constituent of a secured network. Owing to the complexities of network data, one of the challenges with modern detection systems is false alarm and inaccuracy (Zhu et al., 2021).

In this research, an improved algorithm is proposed to efficiently detect intruders on applications deployed over computer networks. The performance evaluation is based on a real-world dataset using several metrics, such as accuracy, precision, recall and F1 score. The findings of this research can contribute

Kasu Journal of Computer Science Vol. 1 No. 3 [September, 2024], pp. 489-496
https://doi.org//10.47514/kjcs/2024.1.3.006
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

to the development of more efficient and accurate intrusion detection systems, which fortifies the security of computer networks.

2. Related Works

Modern intrusion detection systems have high false-positive rates, low detection rates, slow processing speeds, and enormous feature dimensions. Consequently, a multi-module integrated intrusion detection system was developed for high-dimensional imbalanced data (Cui et al., 2023). However, network data's high dimension, complexity and imbalance frustrate intrusion detection process.

A deep learning-based network intrusion detection system was developed that works on data driven network based on dimensionality reduction strategies (Yadav and Kalpana, 2022; Abdulhammed et al., 2019) but this approach is vulnerable to compromise as network attacks increases.

Intrusion detection systems must reduce false alarms, detect unknown threats and improve detection. Subsequently, IoT intrusion detection system was proposed (Zheng et al., 2020). Smart homes and intelligent transportation depend on the IoT. The IoT is vulnerable to hostile threats and current security solutions are inadequate. Intrusion detection protects IoT devices from security breaches but this conventional intrusion detection models lack speed and accuracy.

A dimensionality reduction approach was proposed to improve intrusion detection (Elkassabi et al., 2020). Detection accuracy was improved but too many unrelated attributes impeded the detection process. Numerous researchers have utilized machine learning algorithms to classify network traffic as "harmful" or "regular" and minimize data dimensionality for precision. These systems still suffer high false alarm.

Machine learning algorithms were suggested for big data intrusion detection (Othman et al., 2018) but owing to network's massive data volume, detection accuracy is an issue.

An Intrusion Detection System (IDS) detects attacks in real time by analyzing packets (Almusallam et al., 2017). High-velocity and dimensional data streams complicate IDS implementation. Dimensionality is caused by dynamic data streams, limited memory, high complexity and huge data. Therefore, the main challenge here is to develop an IDS that can work efficiently on data-driven network. To build a reliable data-driven IDS, researchers and practitioners must improve detection accuracy for multi streams data volumes.

Principal Component Analysis (PCA) have been used to minimize network intrusion (Vasan & Surendiran, 2016) based on network traffic analysis. Principal component analysis reduces dimensionality by measuring its reduction ratio, identifying the appropriate number of principal components and assessing the effect of noisy data. However, this approach suffers accuracy problems.

3. Methodology

The algorithm was implemented in the Jupyter Notebook platform. The study used the panda’s library to read the dataset, which was converted to '.csv' format. The dataset's statistical information, such as mean, percentile, and standard deviation, was calculated using the describe function. The dataset used in this research was obtained from the Canadian Institute for Cyber security (CIC), specifically from the CIC Intrusion Detection Evaluation Dataset (CICIDS2017).

The CICIDS2017 dataset is a benchmark dataset for evaluating intrusion detection systems and was created by the CIC research team to support cyber security research. The dataset contains a variety of network traffic data, including both benign and malicious traffic. The malicious traffic includes several

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

different types of attacks, such as Denial of Service (DoS), Distributed DoS (DDoS), Brute Force, PortScan, and Botnet. The dataset also includes data on several network protocols, including TCP, UDP, ICMP and HTTP. The dataset is structured into different CSV files, each containing specific information about the network traffic. For example, there are separate files for flow statistics, payload statistics, and aggregated flow statistics. The dataset is very large, with over 2.8 million instances in total.

The CICIDS2017 dataset is freely available for research purposes and can be downloaded from the CIC website. The dataset has been widely used in the evaluation of intrusion detection systems. Its use in this research is important because it provides a real-world dataset that enables the comparison of different combinations of dimensionality reduction techniques and classification algorithms for intrusion detection based on forensic frameworks. The dataset can be downloaded from the following link: https://www.unb.ca/cic/datasets/ids-2017.html. The dataset was preprocessed to remove redundant or irrelevant features which were normalized. The preprocessing step is essential to ensure that the dataset is suitable for analysis.

The training and testing datasets were prepared using standard scalar and label encoder methods, respectively. The training set was used to fit the model and conduct testing to evaluate the performance of the developed model. The training dataset comprised of 34, 924 rows and 38 columns, while the testing dataset contained 23, 643 rows and 44 columns.

3.1. Proposed Algorithm

The proposed algorithm works by identifying a hyperplane in a high-dimensional space that separates data into different classes. The goal of the algorithm is to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class. It finds the new data point in the feature space and assign the most frequent class label among the neighbors to the new data point. The proposed algorithm is effective when the dataset has a clear separation between the classes and works well for datasets with many features and complex decision boundary. This algorithm is improved to enhance the performance of intrusion detection based on forensic frameworks.

Kasu Journal of Computer Science Vol. 1 No. 3 [September, 2024], pp. 489-496
https://doi.org//10.47514/kjcs/2024.1.3.006
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Proposed Algorithm

Given: (x1; y1) ……, (Xm; Ym), $x_i \in X$ $y_i \in Y = \{=1,1\}$
Initialize D_l (i) = l /m
For t=1…T:
1. Train weak classifier using distribution D_t
2. Get weak hypothesis hy: X — {1,1} with error & = Fy, (4, Di(xi)
3. Choose $\alpha_t = {}^{1}/_{2}\log\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
4. Update:

$$D_{t+1}(i) = \frac{Dt(i)}{Z_t} = \begin{cases} e^{-\alpha}t & \text{if instance i is correctly classified} \\ e^{\alpha_t} & \text{if instance i is not correctly classified} \end{cases}$$

where Z_t, is a normalization factor (chosen so that $\sum_{i=1}^{m} D_{t+1} = 1$)
Output the final Hypothesis: H (x) = sign $(\sum_{t=1}^{T} \alpha_t h_t(x))$
5. GenDecTree (Hypothesis: H (x). Features F)
6. If stopping_condition (H (x). F) = true then
7. Leaf= createNode()
8. leafLabel = classify(s)
9. return leaf
10. root = createNode()
11. root.test_condition =findBestSpilt (H (x). F)
12. V= {v / v a possible outcomecfroot.test_condition}
13. Foreach value v € V:
14. H (x)v = {s / root.test condition(s) = v and s € S}:
15. Child = TreeGrowth (H (x), F):
16. Add child as descent of root and label the edge {root →child} as v
17. return root

3.2. Performance Evaluation

The dataset was splitted into training and testing sets to evaluate the performance of the different components of the algorithms. Performance was evaluated using several metrics, including accuracy, precision, recall, and F1 score (Kabir et al., 2023) as follows:

i. **Accuracy/Recall/Precision/F1 Score/Misclassification Rate:** This is measured by the following metrics: The ratio of correctly predicted observations measures the accuracy. where $accuracy = (TP + TN)/All\ Predictions$

ii. **Sensitivity or true positive rate (TPR):** The probability that, the algorithms can correctly predict positive instances (intrusive ones) to total intrusive instances (TPR) is called sensitivity or true positive rate (TPR). $Transient\ Predictive\ Response\ (TPR) = TP/TP + FN$

iii. **Specificity or (TNR):** It implies the possibility that the algorithms can predetermine negative (normal) instances to the total number of normal instances. $Specificity = TN/TN + FP$

iv. **False-positive rate (FPR):** is defined as the percentage of negative instances predicted by the model to be positive. To calculate the FPR, divide the FPR by the FPR. It is the percentage of correctly predicted positive events that are recalled. In which place: $Recall = TP/(FN + TP)$;

v. **False-negative rate (FNR):** It is the fraction of positive instances that are predicted to be in the negative class, which is known as the false-negative rate (FNR). $FNR = FN\ /\ (TP + FN)$.

Kasu Journal of Computer Science Vol. 1 No. 3 [September, 2024], pp. 489-496
https://doi.org//10.47514/kjcs/2024.1.3.006
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4. Results and discussion

To evaluate the performance of the proposed algorithm, the data were divided into training and testing subsets. The selected features and performance of the classifications were evaluated using confusion matrices. The algorithm's overall performance was evaluated using the Receiver Operating Characteristic (ROC) curve.

Table 1 summarizes the performance measures of the proposed algorithm while Table 2 shows performance of the existing algorithm. These results contribute to the understanding of the algorithm's effectiveness in intrusion detection and provide insights into areas for further improvement.

Table 1. Performance Evaluations of Proposed Algorithm

Metrics	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
Accuracy	98.28	98.70	98.08	98.08
Sensitivity	98.94	98.75	98.07	98.23
Specificity	98.34	98.82	98.35	98.62
Precision	98.75	98.93	98.33	98.65
F1 score	98.66	98.93	98.83	98.95

Table 2. Performance Evaluations of Existing Algorithm (Cui et al., 2023)

Metrics	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration
Accuracy	84.60	83.57	82.45	80.91
Precision	85.13	84.33	83.08	81.85
F1 score	83.95	82.89	81.21	78.13

The results of the experiment indicates that, the improved algorithm exhibited promising results with average accuracy rate of 98.89% in comparison to existing algorithm. The accuracy rates achieved in this research are comparable to those reported in recent studies that have used machine learning techniques for intrusion detection on the CICIDS 2017 dataset. The proposed algorithm outperformed some of the previously reported techniques, indicating that the improved algorithm has the potential to demonstrate enviable prowess.

5. Conclusion

In conclusion, this research examined the performance of an improved algorithm with classifiers for intrusion detection using the CICIDS2017 dataset. The results obtained indicate that the proposed algorithm outperformed the existing ones. The findings of this research suggest that, the use of classification technique in intrusion detection can significantly improve the performance of the detection model. Furthermore, implementing the algorithm into a dynamic and real-time detection system is proposed for future research.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

References

Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., &Abuzneid, A. (2019). Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection. *Electronics*, *8*(3), 322. https://doi.org/10.3390/electronics8030322

Almusallam, N. Y., Tari, Z., Bertok, P., &Zomaya, A. Y. (2017). *Dimensionality Reduction for Intrusion Detection Systems in Multi-data Streams—A Review and Proposal of Unsupervised Feature Selection Scheme* (pp. 467–487). https://doi.org/10.1007/978-3-319-46376-6_22

Al-Yaseen, W. L., Othman, Z. A., &Nazri, M. Z. A. (2017). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. *Expert Systems with Applications*, *67*, 296–303. https://doi.org/10.1016/j.eswa.2016.09.041

Binbusayyis, A., &Vaiyapuri, T. (2020). Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection. *Heliyon*, *6*(7), e04262. https://doi.org/10.1016/j.heliyon.2020.e04262

Chawla, M., &Duhan, M. (2018). Levy Flights in Metaheuristics Optimization Algorithms – A Review. *Applied Artificial Intelligence*, *32*(9–10), 802–821. https://doi.org/10.1080/08839514.2018.1508807

Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Cui, J., Zong, L., Xie, J., & Tang, M. (2023). A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. *Applied Intelligence*, *53*(1), 272–288. https://doi.org/10.1007/s10489-022-03361-2

Elkassabi, H., Ashour, M., &Zaki, F. (2020). A Proposed Model for Dimensionality Reduction to Improve the Classification Capability of Intrusion Protection Systems. *International Journal of Network Security & Its Applications*, *12*(4), 17–37. https://doi.org/10.5121/ijnsa.2020.12402

Emary, E., Zawbaa, H. M., &Sharawi, M. (2019). Impact of Lèvy flight on modern meta-heuristic optimizers. *Applied Soft Computing*, *75*, 775–789. https://doi.org/10.1016/j.asoc.2018.11.033

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2020). *Locally Linear Embedding and its Variants: Tutorial and Survey*.

Gholami, R., &Fakhari, N. (2017). Support Vector Machine: Principles, Parameters, and Applications. In *Handbook of Neural Computation* (pp. 515–535). Elsevier. https://doi.org/10.1016/B978-0-12-811318-9.00027-2

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Hsu, C.-Y., Wang, S., &Qiao, Y. (2021). Intrusion detection by machine learning for multimedia platform. *Multimedia Tools and Applications*, *80*(19), 29643–29656. https://doi.org/10.1007/s11042-021-11100-x

Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences*, *80*(5), 973–993. https://doi.org/10.1016/j.jcss.2014.02.005

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, *8*(3), 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Kabir, M. F., Chen, T., & Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, *3*, 100125. https://doi.org/10.1016/j.health.2022.100125

Kaidi, W., Khishe, M., & Mohammadi, M. (2022). Dynamic Levy Flight Chimp Optimization. *Knowledge-Based Systems*, *235*, 107625. https://doi.org/10.1016/j.knosys.2021.107625

Katoch, S., Chauhan, S. S., & Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, *80*(5), 8091–8126. https://doi.org/10.1007/s11042-020-10139-6

Keerthi Vasan, K., &Surendiran, B. (2016). Dimensionality reduction using Principal Component Analysis for network intrusion detection. *Perspectives in Science*, *8*, 510–512. https://doi.org/10.1016/j.pisc.2016.05.010

Khraisat, A., Gondal, I., Vamplew, P., &Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, *2*(1), 20. https://doi.org/10.1186/s42400-019-0038-7

Malik, H., Fatema, N., & Iqbal, A. (2021). Advances in Machine Learning and Data Analytics. In *Intelligent Data-Analytics for Condition Monitoring* (pp. 3–29). Elsevier. https://doi.org/10.1016/B978-0-323-85510-5.00001-6

Miao, J., Yang, T., Sun, L., Fei, X., Niu, L., & Shi, Y. (2022). Graph regularized locally linear embedding for unsupervised feature selection. *Pattern Recognition*, *122*, 108299. https://doi.org/10.1016/j.patcog.2021.108299

Othman, S. M., Ba-Alwi, F. M., Alsohybe, N. T., & Al-Hashida, A. Y. (2018). Intrusion detection model using machine learning algorithm on Big Data environment. *Journal of Big Data*, *5*(1), 34. https://doi.org/10.1186/s40537-018-0145-4

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Satilmiş, H., Akleylek, S., & Tok, Z. Y. (2024). A Systematic Literature Review on Host-Based Intrusion Detection Systems. Ieee Access, 12, 27237-27266.

Sikos, L. F. (2020). Packet analysis for network forensics: A comprehensive survey. *Forensic Science International: Digital Investigation*, *32*, 200892. https://doi.org/10.1016/j.fsidi.2019.200892

Srikanth Yadav, M., & Kalpana, R. (2022). *Effective Dimensionality Reduction Techniques for Network Intrusion Detection System Based on Deep Learning* (pp. 507–516). https://doi.org/10.1007/978-981-16-6460-1_39

Zhao, S., Li, W., Zia, T., &Zomaya, A. Y. (2017). A Dimension Reduction Model and Classifier for Anomaly-Based Intrusion Detection in Internet of Things. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 836–843. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.141

Zheng, D., Hong, Z., Wang, N., & Chen, P. (2020). An Improved LDA-Based ELM Classification for Intrusion Detection Algorithm in IoT Application. *Sensors*, *20*(6), 1706. https://doi.org/10.3390/s20061706

Zoppis, I., Mauri, G., &Dondi, R. (2019). Kernel Methods: Support Vector Machines. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 503–510). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20342-7