iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68447

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Intrusion Detection System: Comparative Analysis of Supervised and Unsupervised Techniques

Anshita Singh[1], Manish Choudhary[2], Abhishek Singh[3]
[1]*Computer Science & Engineering, BSSITM, Lucknow, India*
[2]*Mechanical Engineering, Central Institute of Petrochemicals Engineering and Technology, Lucknow, India*
[3]*Mechanical Engineering, BSSITM, Lucknow, India*

Abstract: ***With advancements in technology, the rapid growth in cybercrimes poses crucial challenges to maintaining the security and integrity of computer networks. Signature-based techniques and predefined rules are the traditional methods for Intrusion Detection Systems, which are inadequate for handling emerging cybercrimes. This paper presents a comparative analysis of supervised and unsupervised machine learning techniques in Intrusion Detection Systems. Various machine learning models, such as supervised and unsupervised learning, are used to review the limitations of traditional IDS approaches. Overcoming these challenges in conventional Intrusion Detection Systems is the key concept behind this paper. In the supervised learning method we have used Support Vector Machine, Decision Tree and Random Forest and for unsupervised learning algorithm in intrusion detection, we use Principal Component Analysis, K-Means method and DBSCAN. This work discusses the implications of adopting machine learning in intrusion detection and suggests potential areas for future research, such as the integration of deep learning techniques and the development of an adaptive intrusion detection system that evolves with emerging threats. The outcomes are meant to help in the development of more sophisticated and effective intrusion detection systems that are capable of handling novel cyber-attacks.***
Keywords: ***Anomaly detection, cyber security, deep learning, machine learning, network security***

I. INTRODUCTION

In the present era of digitalization, as the world becomes more interconnected, cyber-attacks have become a prominent problem. Internet has bridged the gap by connecting people across the globe but it also jeopardizes users' personal data [1]. Data protection is of paramount importance in contemporary world [2].Organizations can spend millions of dollars on the most secure servers, but it takes a hacker to ruin all the goodwill between them. Conventional mechanisms for handling such crimes such as using firewalls and antivirus software are inadequate against modern cyber-attacks, which exploit system vulnerabilities and fail to maintain confidentiality, integrity, and availability[3].To prevent these malicious attacks many automated security systems have been developed. Machine learning, a subset of Artificial Intelligence, uses various algorithms on trained datasets to enable the automatic detection and respond to a wide range of cyber threats[4]. Intrusion detection systems (IDS) are methods of monitoring network traffic, identifying potentially malicious activities, and safeguarding information systems[5]. IDS technologies occur in many different kinds of forms, including host-based, network-based, and wireless. They all offer basically distinct capabilities for collecting information, recording, detection, and prevention. Additionally, every method offers features including improved precision or efficiency in identifying specific occurrences[6].

Signature-based and anomaly-based systems are traditional approaches of intrusion detection [7].To locate known intrusions signature-based method relies on already specified attack signatures, but they struggle to identify novel or zero-day attacks in contrast the key benefit of anomaly-based detection techniques is their ability to detect previously unrecorded intrusions[8].Any deviation from normal behavior is detected as intrusion. Their potential to identify zero-day intrusions makes them attractive[9].One of the advantages of using this method is that it facilitates the detection of hackers who use novel tactics to penetrate the system [10].Anomaly based systems for intrusion detection primarily consist of - machine learning, statistical and knowledge based [11]. In this review paper we discuss about Machine Learning (ML) methods, a subset of Artificial intelligence that can identify patterns based on historical data, allowing them to learn and make decisions in real time, improving threat detection. Machine learning-based detection techniques for intrusion evaluate massive quantities of network traffic data with greater precision than conventional techniques, enabling them to figure out difference between malicious and legitimate activity. Various types of supervised and unsupervised learning are the two methods used in this review work.

Proposed supervised learning relies on labeled datasets in which system is trained on known instances of malicious behavior whereas unsupervised learning identifies patterns in unlabeled data. The intention of this study is to find improved real-time detection methods by utilizing Machine Learning (ML) strategies to recognize both known and unknown intrusions. This review paper discusses key challenges such as data imbalance, false positives, and the need for real-time processing, along with emerging trends such as federated learning and adversarial defenses. Through its investigation into the current state of research in this area, it is hoped that this paper can provide insights on where to take ML-based intrusion detection systems next and highlight that innovation needs to be ongoing when combating cyber threats.

II. METHODOLOGY

A. *Supervised Learning Methods*

Labeled data is essential for supervised learning or classification in order to train a model for detection[12]. The supervised learning methods discussed below-

1) Support Vector Machine (SVM)

Support Vector Machine is a method based on supervised machine learning and is frequently applied in classification and regression problems. SVM plots training vectors in high-dimensional feature space and assign a class to every single element [13]. It operates by recognizing a hyperplane in a high-dimensional space which most effectively divides data points from distinct classes. The objective is to attain the best possible margin, which is the distance between the hyperplane and the closest data points in each class. A higher margin guarantees better model generalization[14].SVMs can produce good results with small training samples, as it determines the hyperplane of separation using only a few support vectors. But SVMs are noise sensitive across the hyperplane [15].With limited training sets, SVMs are capaable of producing satisfactory results[16].

2) Decision Trees in Intrusion Detection Systems

A commonly used hierarchical approach is the decision tree, where every node within the tree implies a selection depending on a feature, every connected branch indicates a decision outcome, and a leaf node depicts a class label [17]. Decision trees have the capacity to assess data and detect important attributes of networks that point to fraudulent activity[18]. It is interpretable and flexible machine learning paradigm in which data is divided into subsets based on the feature values at each internal node, making them highly comprehend and appropriate for classification and regression tasks[19].

The decision tree classification algorithm's ease of use is one of the primary advantages since it does not need the user to have a lot of prior knowledge to comprehend. The technique can efficiently learn from the training instances as long as they are provided with characteristics and conclusions [20]. If it comes to extracting features and rules, it provides several advantages. Decision tree algorithms are therefore used in the intrusion detection field [21].Large datasets are an ideal match for decision trees. Decision trees are beneficial for real-time intrusion detection because of their outstanding efficiency [22].

3) Random Forest Method

This approach is based on supervised learning technique for intrusion detection because of its ability to handle complex datasets and classify network behavior effectively. Random Forest is an ideal supervised learning strategy that can develop a model that can forecast the classification outcomes of a specific sample type belonging to a given dataset based on its distinct features and classification results[23].An ensemble approach called Random forest generates predictions using the outcomes of several decision trees[24].After the forest has been constructed, another item that requires to be classified is placed at each tree in the forest for grouping. Beginning with the root, the decision tree approach analyzes each node in order to determine the most suitable division among all available features, whereas in the random forest method, just an arbitrary number of features are chosen for evaluation [25]. Compared with other conventional classification techniques, Random Forest exhibits minimal errors in classification [26].One of the most significant algorithms for classification is the Random Forest algorithm, which effectively categorizes huge quantities of data [27].

B. *Unsupervised Learning*

Unsupervised learning approach aim to identify irregularities in the system or network's behavior, which could signify potential intrusions. In unsupervised learning, the training data lacks labels, and the model independently explores the data to uncover inherent patterns [28]. Here are some popular unsupervised methods used in IDS

1) K-Means

K-Means is a well-known unsupervised learning algorithm used for clustering[29]. In the context of Intrusion Detection Systems (IDS), K-Means can group network traffic data into clusters, helping to identify potential intrusions by flagging abnormal or unusual data points that deviate from the normal clusters[30]. The K-means method is often implemented in time series data for the identification of patterns [31].Data points that are away from the centroids of groups or do not belong to any cluster are used to identify threats[32]. The approach successfully identifies anomalous behavior with a high degree of accuracy while being computationally efficient. By fragmenting "n" data points into "k" clusters, the technique known as K-Means assigns each point to the cluster with the mean that is nearest to it. Since it is a distance-based clustering approach, it lacks the ability to calculate distances for each pair of records that could possibly exist [33].Once all points are allocated, the first step concludes with an initial grouping. Next, K new centroids are computed for the clusters formed, ensuring that the intra-cluster distance to the centroid is minimized. This process is repeated iteratively until the centroids remain unchanged [34]. The primary aim of the K-Means clustering technique is to segment and classify data into attack and normal instances.

2) Density-Based Spatial Clustering Of Applications With Noise (DBSCAN)

DBSCAN is a powerful unsupervised learning algorithm widely used in intrusion detection systems (IDS). It is especially effective for anomaly detection because it identifies dense regions in data and labels points in low-density regions as outliers, which often correspond to intrusions or anomalous activities[35]. The key idea is that the majority of the data is regular, and the normal data will be clustered into a higher-density cluster, whereas the intrusion data will be limited and quite different from the normal data. So, using clustering, invasion data would be a low-density cluster [36].In sample data, the DBSCAN algorithm mainly recognizes groups as a dense area of events separated by low-density zones [37].

3) Principle Component Analysis

Another dimensionality reduction technique that could be extremely useful in intrusion detection systems (IDS) is principal component analysis (PCA)[38].PCA is achieved by projecting the most pertinent and valuable attributes into a lower dimensional subspace and removing the features that are not as significant in the higher dimensional space which has the highest computational cost [39].The underlying concept is to reduce an excessive number of variables into a smaller number of independent variables by recognizing a small number of orthogonal linear combinations of the original variables with the largest variance[40].The quality of the dataset will become better since it might have suitable properties[41].

III. RESULT & DISCUSSION

This review paper investigates and compares the performance of unsupervised learning techniques—such as K-Means, Principal Component Analysis (PCA), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)—and supervised learning algorithms—such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF)—in identifying intrusions. The key points on the following parameters such as accuracy, precision, determining efficacy, and flexibility in response to new threats are key points discussed in this paper. Here are some key points:

A. Supervised Learning-Based IDS Performance

1) As SVM has a strong mathematical framework for classification, it can distinguish between malicious and legitimate information. It performs well in multidimensional spaces, but due to its computational complexity, it has difficulties with massive datasets.

2) In structured datasets, the Decision tree provides swift and precise intrusion detection with an outstanding efficiency; however, it is vulnerable to overfitting, especially when trained on complex datasets with many attributes.

3) In the Random Forest, the overall precision and generalization are higher than Support Vector Machine and Decision Tree because it utilizes ensemble learning to decrease variance. It is less prone to overfit than decision tree, but it needs a greater amount of processing power.

B. Unsupervised Learning-Based IDS Performance

1) K-Means clustering works well in classifying identical attack patterns, but it has fixed cluster assumptions that may not align with the broad spectrum of attacks occurring in the real world. This approach is sensitive to initial cluster assignments, which may lead to the misclassification of uncommon attacks.

2) Principal component analysis effectively preserves variance in intrusion data while minimizing the dimensionality of the data. It improves classification efficiency when merged with additional machine learning models; however, it has issues with non-linear attack patterns.
3) DBSCAN efficiently recognizes deviations and anomalies, such as unidentified cyber threats. It does not require predefined cluster numbers, making it more flexible than K-Means. However, it can be computationally expensive for large datasets and performs poorly in high-dimensional spaces

IV. CONCLUSION

The review work aimed to explore and compare the effectiveness of supervised and unsupervised machine learning techniques in Intrusion Detection Systems (IDS). By analyzing different approaches, this study has shown that both supervised and unsupervised methods offer unique strengths. Support Vector Machines (SVM) are supervised learning techniques that perform well with small labeled datasets, while Decision Tree and Random Forest methods produce better results with large training samples Although these supervised learning methods are effective at recognizing known threats, they might not be as efficient against zero-day attacks. In contrast, unsupervised methods like K-Means, Principal Component Analysis (PCA), and DBSCAN are effective for detecting unusual patterns and anomalies that may indicate unknown or evolving threats. These methods do not require labeled data, making them particularly useful in dynamic network environments where labeling every instance is impractical.

However, each technique also presents its own limitations. The reliability of supervised algorithms decreases significantly if unknown attacks are present in the test data, as supervised methods rely mostly on labeled datasets. Unsupervised methods, while adaptable, often yield higher false positive rates, necessitating additional filtering or combination with other methods to enhance accuracy. Overall, this comparative analysis underscores the importance of selecting effective machine learning algorithms based on the specific requirements and limitations of the security environment. With the goal to cope with the increasing requirements in modern cyber security, future research may focus on either enhancing hybrid models or developing new techniques that mitigate the limitations the negative of strictly supervised or unsupervised approaches.

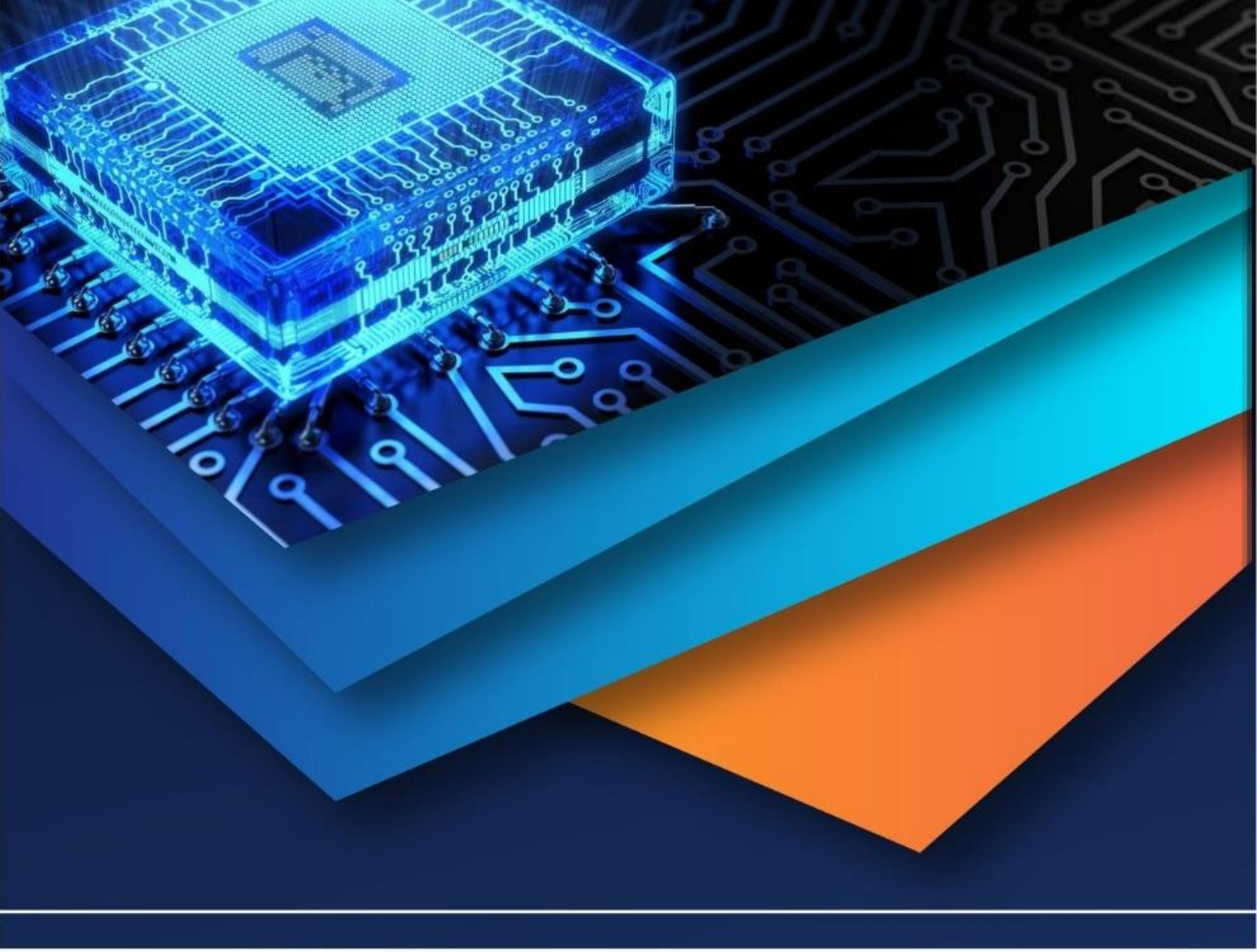
Conflict of Interest: There are no conflicts of Interest among authors.

Acknowledgement: The authors are thankful to their research places for providing the support for this work.

REFERENCES

[1] Cyber Security Threats and Countermeasures in Digital Age. Journal of Applied Science and Engineering, vol. 4, no. 1 (2024): 1–20. https://doi.org/10.54060/a2zjournals.jase.42.
[2] Saeed, S., S. A. Altamimi, N. A. Alkayyal, E. Alshehri, and D. A. Alabbad. “Digital Transformation and Cybersecurity Challenges for Businesses Resilience: Issues and Recommendations.” Sensors 23, no. 15 (2023): 1–20. https://doi.org/10.3390/s23156666.
[3] Neupane, K., R. Haddad, and L. Chen. “Next Generation Firewall for Network Security: A Survey.” Conference Proceedings - IEEE SOUTHEASTCON (2018): 1–6. https://doi.org/10.1109/SECON.2018.8478973.
[4] Sarker, I. H. “Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects.” Annals of Data Science 10, no. 6 (2023): 1473–1498. https://doi.org/10.1007/s40745-022-00444-2.
[5] Lunt, T. F. “A Survey of Intrusion Detection Techniques.” Computers & Security 12, no. 4 (1993): 405–418. https://doi.org/10.1016/0167-4048(93)90029-5.
[6] Ozkan-Okay, M., R. Samet, O. Aslan, and D. Gupta. “A Comprehensive Systematic Literature Review on Intrusion Detection Systems.” IEEE Access 9 (2021): 157727–157760. https://doi.org/10.1109/ACCESS.2021.3129336.
[7] Rama Devi, R., and M. Abualkibash. “Intrusion Detection System Classification Using Different Machine Learning Algorithms on KDD-99 and NSL-KDD Datasets - A Review Paper.” International Journal of Computer Science and Information Technology 11, no. 03 (2019): 65–80. https://doi.org/10.5121/ijcsit.2019.11306.
[8] García-Teodoro, P., J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez. “Anomaly-Based Network Intrusion Detection: Techniques, Systems and Challenges.” Computers & Security 28, no. 1–2 (2009): 18–28. https://doi.org/10.1016/j.cose.2008.08.003.
[9] Buczak, A. L., and E. Guven. “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection.” IEEE Communications Surveys & Tutorials 18, no. 2 (2016): 1153–1176. https://doi.org/10.1109/COMST.2015.2494502.
[10] Einy, S., C. Oz, and Y. D. Navaei. “The Anomaly- And Signature-Based IDS for Network Security Using Hybrid Inference Systems.” Mathematical Problems in Engineering 2021 (2021). https://doi.org/10.1155/2021/6639714.
[11] Samrin, R., and D. Vasumathi. “Review on Anomaly Based Network Intrusion Detection System.” International Conference on Electrical, Electronics, Communications, and Computing Technologies Optimization Techniques (ICEECCOT) 2017, vol. 2018 (2017): 141–147. https://doi.org/10.1109/ICEECCOT.2017.8284655.
[12] Abdallah, E. E., W. Eleisah, and A. F. Otoom. “Intrusion Detection Systems Using Supervised Machine Learning Techniques: A Survey.” Procedia Computer Science 201, no. C (2022): 205–212. https://doi.org/10.1016/j.procs.2022.03.029.
[13] Mukkamala, S., G. Janoski, and A. Sung. “Intrusion Detection Using Neural Networks and Support Vector Machines.” Proceedings of the International Joint Conference on Neural Networks 2 (2002): 1702–1707. https://doi.org/10.1109/ijcnn.2002.1007774.

[14] Ahmad, I., M. Basheri, M. J. Iqbal, and A. Rahim. "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection." IEEE Access 6, no. c (2018): 33789–33795. https://doi.org/10.1109/ACCESS.2018.2841987.
[15] Johnson. "The Journal of Computing Sciences in Colleges." Journal of Computing Sciences in Colleges 34, no. 3 (2019): 1–120.
[16] Liu, H., and B. Lang. "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey." MDPI AG (2019): https://doi.org/10.3390/app9204396.
[17] Parameswari, D., and V. Khanaa. "Intrusion Detection System Using Modified J48 Decision Tree Algorithm." Journal of Critical Reviews 7, no. 4 (2020): 730–734. https://doi.org/10.31838/jcr.07.04.135.
[18] Rai, K., M. S. Devi, and A. Guleria. "Decision Tree Based Algorithm for Intrusion Detection." International Journal of Advanced Networking Applications 7, no. 4 (2016): 2828–2834. https://www.researchgate.net/publication/298175900.
[19] Ahmad, Z., A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad. "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches." Transactions on Emerging Telecommunications Technologies 32, no. 1 (2021): https://doi.org/10.1002/ett.4150.
[20] Wang, J., Q. Yang, and D. Ren. "An Intrusion Detection Algorithm Based on Decision Tree Technology." Proceedings of the 2009 Asia-Pacific Conference on Information Processing 2 (2009): 333–335. https://doi.org/10.1109/APCIP.2009.218.
[21] Relan, N. G., and P. G. Student. "13=6-TREE.pdf." (2015): 3–7.
[22] Peddabachigari, S., A. Abraham, and J. Thomas. "Intrusion Detection Systems Using Decision Trees and Support Vector Machines." (2004).
[23] Liu, L., P. Wang, J. Lin, and L. Liu. "Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning." IEEE Access 9 (2021): 7550–7563. https://doi.org/10.1109/ACCESS.2020.3048198.
[24] Resende, P. A. A., and A. C. Drummond. "A Survey of Random Forest-Based Methods for Intrusion Detection Systems." ACM Computing Surveys 51, no. 3 (2018): https://doi.org/10.1145/3178582.
[25] Soheily-Khah, S., P. F. Marteau, and N. Bechet. "Intrusion Detection in Network Systems through Hybrid Supervised and Unsupervised Machine Learning Process: A Case Study on the ISCX Dataset." Proceedings of the 2018 1st International Conference on Data Intelligence and Security (2018): 219–226. https://doi.org/10.1109/ICDIS.2018.00043.
[26] Farnaaz, N., and M. A. Jabbar. "Random Forest Modeling for Network Intrusion Detection System." Procedia Computer Science 89 (2016): 213–217. https://doi.org/10.1016/j.procs.2016.06.047.
[27] Aung, Y. Y., and M. M. Min. "An Analysis of Random Forest Algorithm-Based Network Intrusion Detection System." Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (2017): 127–132. https://doi.org/10.1109/SNPD.2017.8022711.
[28] Ahanger, A. S., S. M. Khan, and F. Masoodi. "An Effective Intrusion Detection System Using Supervised Machine Learning Techniques." Proceedings of the 5th International Conference on Computer Methodologies and Communications (2021): 1639–1644. https://doi.org/10.1109/ICCMC51019.2021.9418291.
[29] Aung, Y. Y., and M. M. Min. "An Analysis of K-means Algorithm-Based Network Intrusion Detection System." Advances in Science, Technology and Engineering Systems 3, no. 1 (2018): 496–501. https://doi.org/10.25046/aj030160.
[30] Khaddor, M. A., and B. Al-Khattib. "Intrusion Detection Systems Using K-Means and Random Forest Algorithms." International Journal of Science and Engineering Research 11, no. 9 (2020): 217–224. http://www.ijser.org.
[31] Saranya, T., S. Sridevi, C. Deisy, T. D. Chung, and M. K. A. A. Khan. "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review." Procedia Computer Science 171 (2020): 1251–1260. https://doi.org/10.1016/j.procs.2020.04.133.
[32] Jianliang, M., S. Haikun, and B. Ling. "The Application of Intrusion Detection Based on K-means Cluster Algorithm." Proceedings of the 2009 International Forum on Information Technology and Applications 1 (2009): 150–152. https://doi.org/10.1109/IFITA.2009.34.
[33] Dumoulin, J., et al. "UNICITY: A Depth Maps Database for People Detection in Security Airlocks." Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance (2018): https://doi.org/10.1109/AVSS.2018.8639152.
[34] Laskov, P., D. Patrick, and C. Sch. "Learning Intrusion Detection: Supervised or Unsupervised?" (2014): 50–57.
[35] Deng, D. "Research on Anomaly Detection Method Based on DBSCAN Clustering Algorithm." Proceedings of the 2020 5th International Conference on Information Science, Computing Technology, and Transportation (2020): 439–442. https://doi.org/10.1109/ISCTT51595.2020.00083.
[36] Li, X. Y., G. H. Gao, and J. X. Sun. "A New Intrusion Detection Method Based on Improved DBSCAN." Proceedings of the 2010 WASE International Conference on Information Engineering 2 (2010): 117–120. https://doi.org/10.1109/ICIE.2010.123.
[37] Jain, P., M. S. Bajpai, and R. Pamula. "A Modified DBSCAN Algorithm for Anomaly Detection in Time-Series Data with Seasonality." Journal of Data Science 19, no. 1 (2022): 23–28.
[38] Abdulhammed, R., H. Musafer, A. Alessa, M. Faezipour, and A. Abuzneid. "Features Dimensionality Reduction Approaches for Machine Learning-Based Network Intrusion Detection." Electronics 8, no. 3 (2019): https://doi.org/10.3390/electronics8030322.
[39] Praneeth, N. S. K. H., N. M. Varma, and R. R. Naik. "Principal Component Analysis-Based Intrusion Detection System Using Support Vector Machine." 2016 IEEE International Conference on Recent Trends in Electronics, Information, and Communication Technologies (2017): 1344–1350. https://doi.org/10.1109/RTEICT.2016.7808050.
[40] Wang, W., and R. Battiti. "Identifying Intrusions in Computer Networks with Principal Component Analysis." Proceedings of the First International Conference on Availability, Reliability and Security (2006): 272–279. https://doi.org/10.1109/ARES.2006.73.
[41] Waskle, S., L. Parashar, and U. Singh. "Intrusion Detection System Using PCA with Random Forest Approach." Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC) (2020): 803–808. https://doi.org/10.1109/ICESC48915.2020.9155656.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY