JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024
Research Article
https://doi.org/10.56979/801/2024

Optimized AI-Driven Intrusion Detection in WSNs: A Semi-Supervised Learning Paradigm

Syed Shahid Abbas[1], Salahuddin[1*], Abdul Manan Razzaq[1], Mubashar Hussain[2], Meiraj Aslam[1], Prince Hamza Shafique[1], and Muhammad Asif Nadeem[3]

[1]Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.
[2]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan.
[3]Department of English, Institute of Southern Punjab, Multan, Punjab.
*Corresponding Author: Salahuddin. Email: msalahuddin8612@gmail.com

Received: March 29, 2024 **Accepted:** September 28, 2024

Abstract: In this study, we have developed an advanced semi-supervised learning model specifically designed to identify four distinct types of attacks in Wireless Sensor Networks (WSNs): Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). Our model leverages the combined advantages of supervised and unsupervised learning approaches, employing a Support Vector Machine (SVM) for the supervised aspect and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for the unsupervised component. We rigorously tested and validated our model using the NSL-KDD dataset, which highlighted its strong performance metrics, including accuracy and F1-score. Additionally, our research investigated the sensitivity of DBSCAN parameters and their effects on model accuracy, underscoring the importance of precise parameter tuning to achieve optimal results. A notable advantage of our semi-supervised approach is its capacity to manage large amounts of unlabeled data effectively, a challenge that purely supervised or unsupervised methods often face independently. By efficiently utilizing labeled data and integrating clustering techniques, our model shows improved accuracy and effectiveness in detecting intrusions within WSNs. Overall, this research advances the field of intrusion detection in WSNs by introducing a practical and effective semi-supervised learning framework. This framework enhances detection performance across various attack types and provides valuable insights into optimizing model performance through parameter sensitivity analysis and strategic dataset use.

Keywords: Intrusion Detection System; Supervised Learning; User to Root; Wireless Sensor Networks; Dataset.

1. Introduction

In this work, we study the deployment and operation of wireless sensor networks (WSNs), which consist of many sensor nodes distributed over an area to collect and monitor data from multiple locations. Each sensor node has a microprocessor and communicates through a central system connected to the Internet.

Wireless sensor networks have extended widespread focusing due to their presentations in many fields, such as environmental monitoring, military, and healthcare. Despite increasing adoption, security remains a major concern due to limited resources that hinder the use of traditional security measures such as encryption and authentication. WSNs are vulnerable to a variety of security threats, including node compromise, malicious attacks, denial of service attacks, and other threats. Here's how: Detect and mitigate these threats. Two main groupings of IDS are false IDS (MIDS), which identify attack patterns based on similar patterns, and anomalous IDS (AIDS), which detect deviations from normal behavior. While MIDS cannot detect new types of attacks, AIDS generally has a lower index.

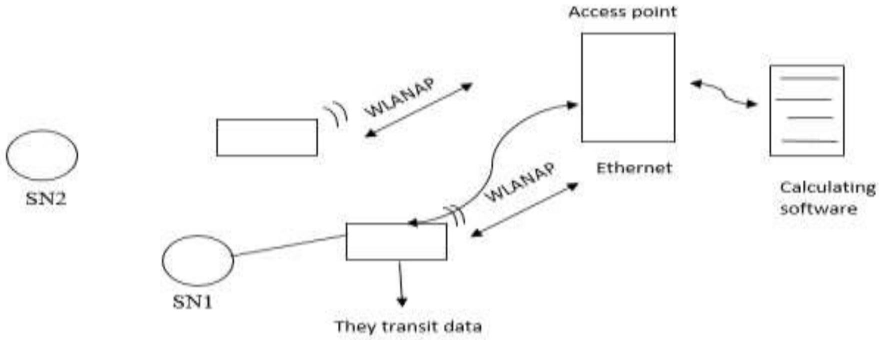

Figure 1. Components of WSN

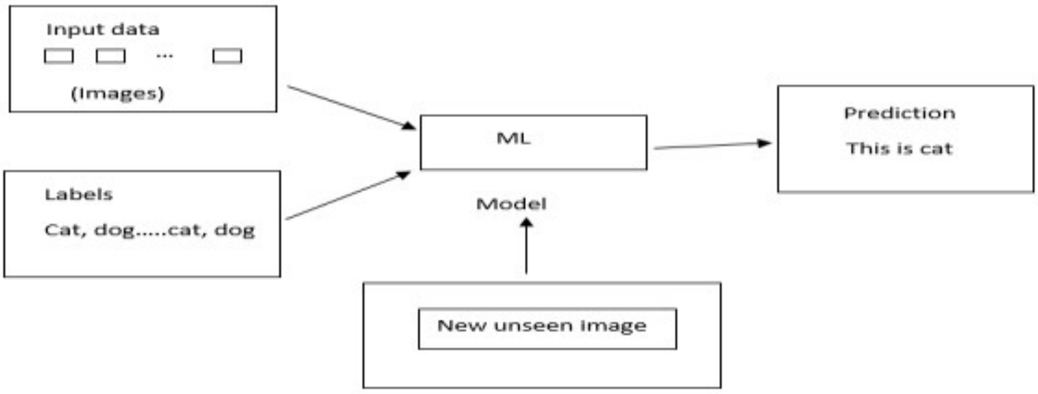

Figure 2. Supervised learning

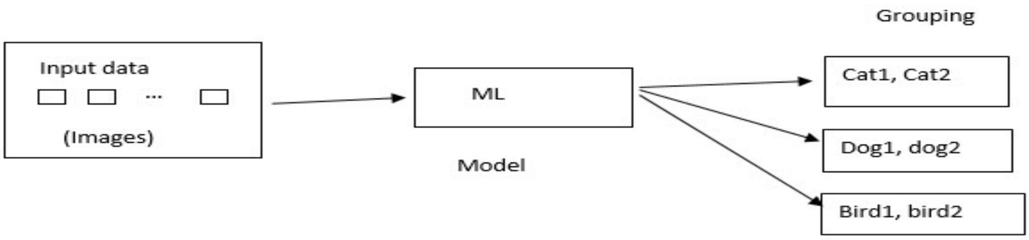

Figure 3. Unsupervised learning

To address these security issues, innumerable machine learning techniques (especially supervised learning, unsupervised learning, and semi-supervised learning) are being studied to improve detection systems (IDS) in WSNs. Supervised learning uses local data to train models to make predictions about new ideas. In contrast, unsupervised learning analyzes unlabeled data to reveal underlying patterns without making any prior decisions. Semi-supervised learning provides an unbiased model using both labeled and unlabeled data, which makes it especially useful in situations where a large amount of labeled data is not available. Ability to solve complex problems. Techniques such as joint analysis, such as support vector machine, decision tree, DBSCAN and K-Means are quite important in developing WSN intrusion detection model. They can provide the following benefits: They are expected to improve the security and reliability of wireless sensor networks, paving mode for their continuous deployment and widespread use.

2. Materials and Methods

This research is designed to develop a semi-supervised learning-based intrusion detection system precisely intended for WSN. Core goal is to improve the detection accuracy while reducing body wear and saving energy. The main objectives include:

2.1. Data collection and prioritization:

Usage NSL-KDD dataset for training and validation of IDS to ensure its reliability and performance in practical applications. Test different ways to extract and select important data for intrusion detection, thereby improving the ability to detect malicious activities. Use IDS with unsupervised learning for unlabeled object matching and supervised learning for model training. This approach aims to strike a balance between accuracy and scalability.

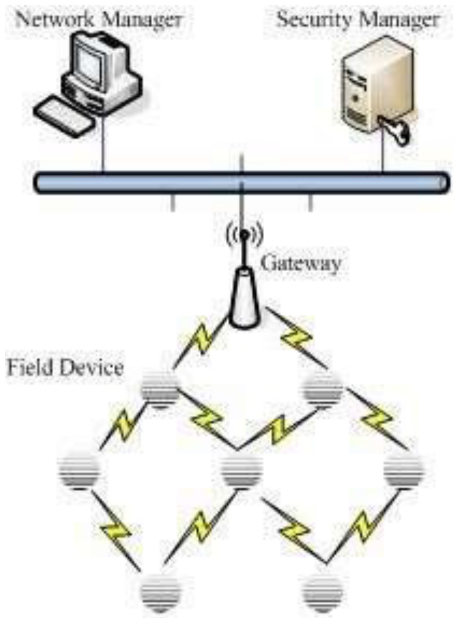

Figure 4. Architecture of WSN

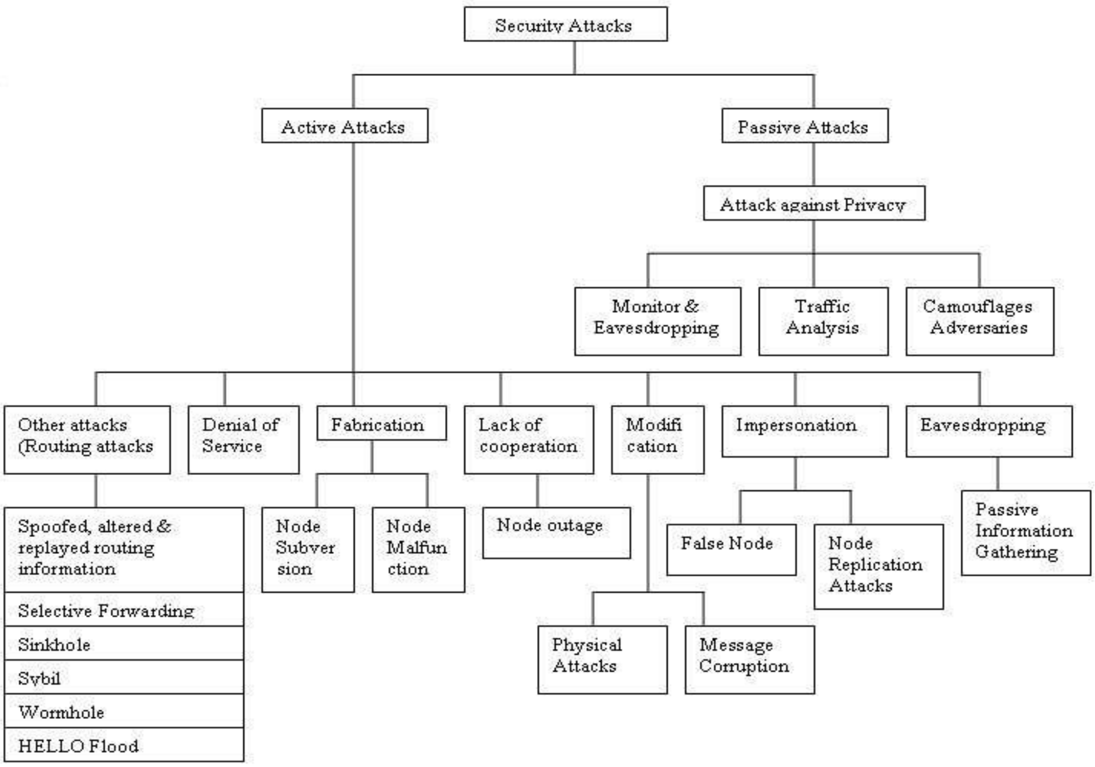

Figure 5. General Classification

2.2. Performance Evaluation

Appraise performance of IDS with respect to detection accuracy, precision, recall, and F1 score to verify its effectiveness in mitigating WSN traffic security threats. Significance of the Research - Monitoring efforts in IDS development overcome the limitations associated with monitoring only or no monitoring.

The model aims to improve the possibility of detecting intrusions without the resources to utilize full data collection, using recorded and unrecorded data. Play an important role in improving technological capabilities in multiple domains:

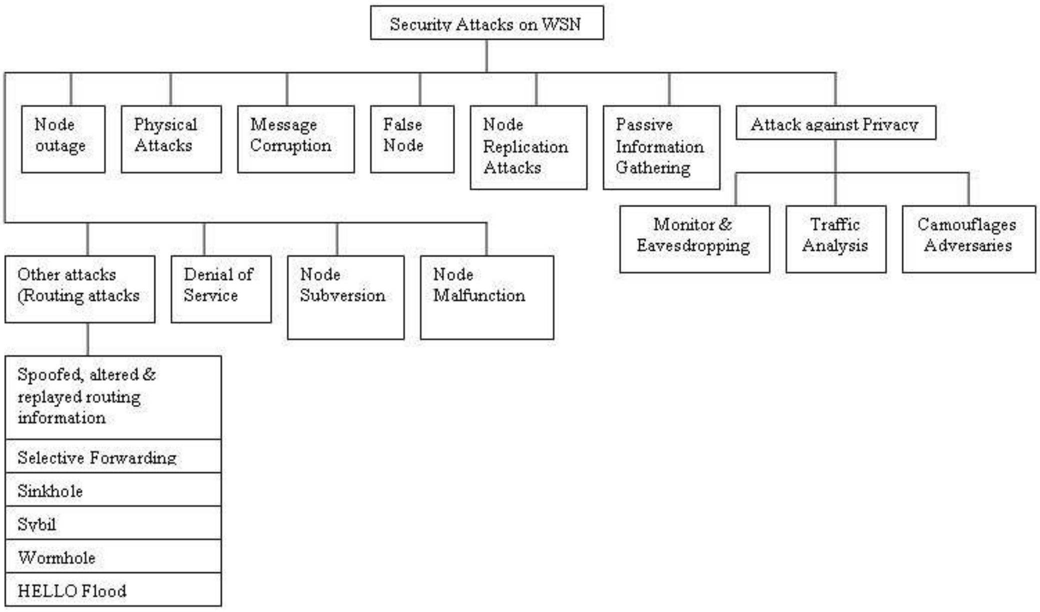

Figure 6. Security Attacks on WSN

2.3. Unsupervised Learning:

Analyze anonymous data to discover hidden patterns like customer behavior based on aggregated data. >- Semi-supervised study: Combine labeled and unlabeled data to come up with a more efficient model, especially in cases where data collection is limited or expensive to obtain.) consists of a network of interconnected sensor nodes designed to collect data remotely:

2.4. Sensor Node:

Consists of various components like sensors, processors and Transceivers, usually powered by batteries or energy harvesting technology.

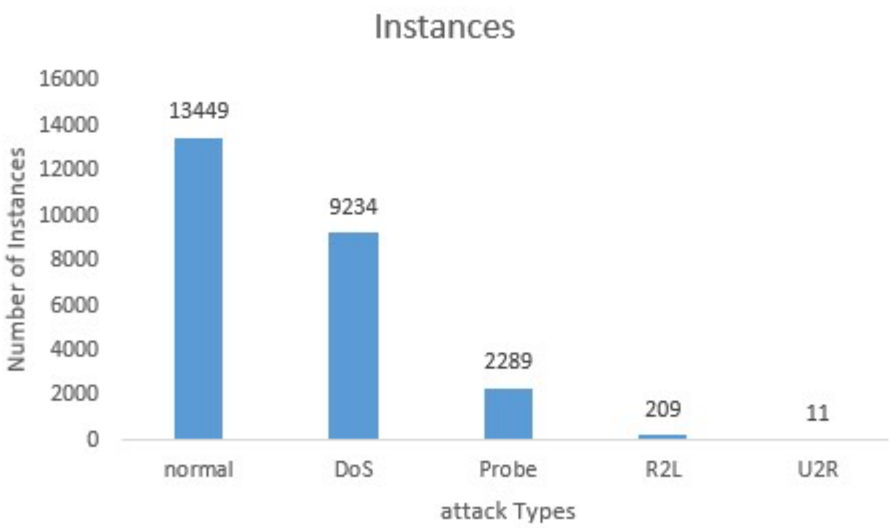

Figure 7. Dataset Instances

Table 1. Attribuites of Dataset

Attribute No.	Attribute Name	Narrative	Sample

1	Duration	Period span for the link	0
2	Protocol_type	Protocol that was employed in the connection construction	Tcp
3	Service	Use of final location networking service	ftp_data
4	Flag	Connection's qualitygood or faulty	SF
5	Src_bytes	Number of data reassigned from origin to recipient in just one linking	491
6	Dst_bytes	Number of data transferred from origin to recipient in just one connection	0
7	Land	If origin and recipient Internet Protocol(IP) addresses and number of ports are same, then this parameter has value of 1 else it possess are value of 0	0
8	Wrong_fragment	Overall extents of improper segments in this link	0
9	Urgent	Quantity of crucial packets in this link imperative packets are those that have vital bit set	0
10	Hot	There are several "hot" indications in data such as reaching system directory , producing programs and executing programs	0

11	Num_failed_logins	The number of unsuccessful attempts to login	0
12	Logged_in	Login situation: 1 if logged in successfully;0 else	0
13	Num_compromised	Numerous “compromised” scenarios	0
14	Root_shell	If root shell is achieved 1 is returned otherwise 0 is returned	0
15	Su_attempted	If “su root “ function attempts to be executed or employed 1 is returned otherwise 0 is returned	0
16	Num_root	Count of “root” visits or activities conducted in the link as root	0
17	Num_file_creations	Entire number of file generation processes carried out through the linking	0
18	Num_shells	Count of shell prompts	0
19	Num_access_files	Numeral of activities performed on admittance resistor files	0
20	Num_outbound_cmds	The total number of outbound commands in single ftp Session	0
21	Is_hot_login	1 if login is on ‘hot’ list of items i.e. root or admin ; otherwise 0	0

22	Is_guest_login	If login is "guest" login 1 is returned otherwise 0 is returned	0
23	Count	In last two seconds total number of connections to exactly same destination host as present connection	2
24	Srv_count	In preceding two seconds, total number of connections to accurately alike amenity port as current connection	2
25	Serror_rate	Fraction of associates in count(23) that have triggered the flag (4) ,s0 ,s1, s2 or s3 out of all connections	0
26	Srv_serror_rate	Fraction of associates in srv_count(24) that have triggered the flag (4) ,s0 ,s1, s2 or s3 out of wholly influences	0
27	Rerror_rate	Fraction of associates with flag(4)REJ that have been activated among the connections aggregated in count(23)	0
28	Srv_rerror_rate	Fraction of associates with flag(4)REJ that have been activated among the connections aggregated in srv_count(24)	0

29	Same_srv_rate	Fraction of associates to identical same service among all connections collected in count (23)	1
30	Diff_srv_rate	Proportions of associates to unlike service amongst connections together in count (23)	0
31	Srv_diff_host_rate	The proportion of connection collected in srv_count(24) that were to various destination machines	0
32	Dst_host_count	Number of connections that have equivalent terminus host IP address	150
33	Dst_host_srv_count	Number of connections consuming the alike port number	25
34	Dst_host_same_srv_rate	Number of connections to alike service that were serene in dst_host_count(32)	25
35	Dst_host_diff_srv_rate	Number of connections to different service that were collected in dst_host_count(32)	0.03
36	Dst_host_same_src_port_rate	Proportion of connections to unchanged origin port that were assembled in dst_host_srv_count(33)	0.17

37	Dst_host_srv_diff_host_rate	The proportion of connections to different origin port that were gathered in dst_host_srv_count(33	0
38	Dst_host_serror_rate	The fraction of connections in dst_host_count(32) that have the flag (4) s0, s1, s2, s3 enabled	0
39	Dst_host_srv_serror_rate	Fraction of associates amongst associates aggregated in dst_host_srv_count(33) that have triggered flag(4) s0, s1, s2 or s3	0
40	Dst_host_rerror_rate	Fraction of associates with flag(4)REJ that have been initiated among the associates aggregated in dst_host_count(32)	0.05
41	Dst_host_srv_rerror_rate	Fraction of associates with flag(4)REJ that have been triggered among the connections aggregated in dst_host_count(33)	0

This study demonstrates how semi-supervised learning can improve IDS capabilities in wireless sensor networks, thus helping to improve network security and reliability for applications such as perimeter protection, travel, and military. In terms of efficiency and effectiveness of access search, it is in line with the current development of machine learning and wireless sensor network technology. Conclusion In our study, we conducted a comprehensive evaluation of the performance indicators (accuracy, F1 score, recall, and precision) of SSDBSCAN link monitoring (SVM) and unsupervised (DBSCAN) procedures. We vary the MinPts and epsilon parameters of DBSCAN to evaluate their effects on network IDS. This paper covers four attack types, including DoS, detection, R2L (remote-to-local), and U2R (user-to-root) distribution and attacks, with a total of 22,638 patients and 29 characteristics. We constructed an overlapping system containing DoS attack clusters after SVM training, clustering, and analysis using SSDBSCAN.

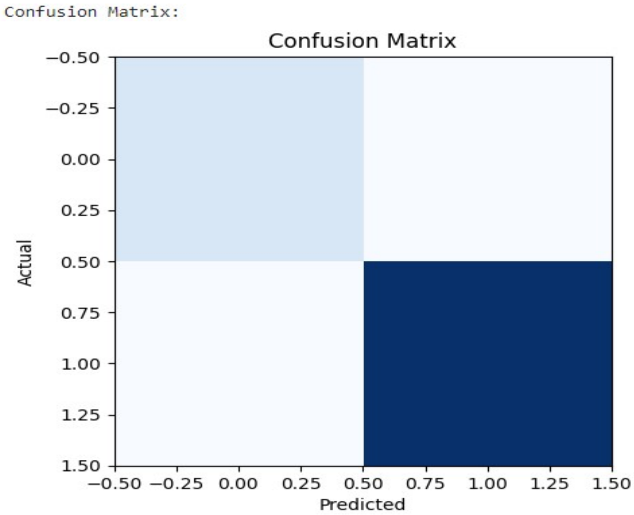

Figure 8. Graph of Confusion Matrix

We experimented with varying MinPts (3, 5, 8, 10) and epsilon (0.2, 0.5, 0.8, 1) values in DBSCAN, resulting in the following performance metrics:

- Comparison of Accuracy with different Eps and MinPts for DoS Dataset:

Similarly, we analyzed the Probe dataset from NSL-KDD, which contains 15,738 instances with the same 29 attributes.

2.5. Confusion Matrix for Probe Dataset:

We evaluated performance metrics for different MinPts and epsilon values.

2.6. Comparison of accuracy of different probe data Eps and MinPts

These tests show the sensitivity of SSDBSCAN performance to measurement parameters. The agreement between minPts and epsilon values can affect the accuracy and precision of detection, which indicates that the access parameter for accessing wireless sensor networks should be carefully selected.

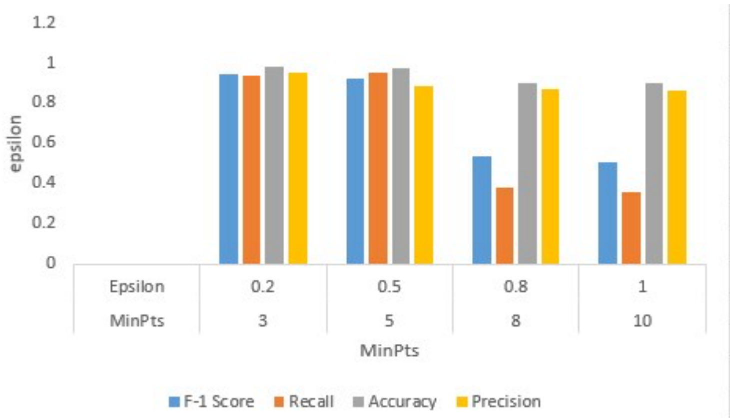

Figure 9. Comparison Graph

Properties of Dataset	Measure
The total number of instances in dataset	13658
The number of features in Dataset	29
Features data types	Nominal, numeric

Figure 10. R2L Dataset Attributes

Table 2. Determining Eps and MinPts in DBSCAN

MinPts	Epsilon	F1-score	Recall	Accuracy	Precision
3	0.2	0.09132	0.04784	0.98542	1.0
5	0.5	0.09132	0.04784	0.98542	1.0
8	0.8	0.09009	0.04784	0.98521	0.7692
10	1	0.0896	0.04784	0.98513	0.71428

3. Discussion

That’s right! This is a better way to write and view content together. Here are examples of tools used:
3.1. Support Vector Machines (SVM)

SVM is an advanced machine learning technique used to classify tasks and can work on both linear and non-linear features. Famous. It works by creating a hyperplane that separates different clusters in a given space. SVM works well with numerical and categorical data. W \cdot x_i + b \geq +1 \) has the form: \ (y_i = +1 \)- \(w \cdot x_i + b \leq -1 \) = \(y_i = -1 \ ) >

Combining these constraints we obtain the following condition:

\ [y_i (w \cdot x_i + b) - 1 \geq 0, \forall i \]

Map* * < br > an object map provides a list of categories. In clustering algorithms, clustering datasets are used to create clusters based on other features. The SVM learning model then determines the appropriate text for this group. Cluster analysis algorithm based on object density. It has two values:

Eps: Defines the maximum radius of the surrounding area. The key points are close to major sites but do not meet the MinPts requirements. Noise is not suitable for certain groups. It then creates clusters by connecting key points in Eps and assigns cluster symbols to the closest points. The algorithm handles noise well and handles different groups and sizes well without any prior sharing. We monitor the performance of the model using the following metrics:

Confusion Matrix: “Clues are Positive (TP), Negative (TN), Factor Negative (FP) and Computed Factor Negative (FN)). The formula is as follows:

\ [\text{Recall} = \frac{TP}{TP + FN} \]”

Precision: “Shows the success of each prediction rig. Nice the important thing is that price and output are equal to these two parameters. The formula is as follows:

\ [\text{F-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \]”

Accuracy: Measurement of the accuracy of the model. The formula is as follows:

“\ [\text{Accuracy} = \frac{TP + TN} {TP + TN + FP + FN} \]”

4. Conclusion

This parameter is also for the operation of the test rig. Large-scale operational models, including semi-supervised models such as SSDBSCAN, are important for finding access to operational systems and evaluating their consequences. The model combines supervised (Support Vector Machine, SVM) and unsupervised methods (DBSCAN) to detect four types of attacks (DoS, Probe, R2L, U2R). We use the NSL-KDD dataset to validate our method and achieve good results in terms of accuracy and F1 score. By changing the parameters in DBSCAN, we observed changes in accuracy, indicating the sensitivity of the model to changes. The balance between clean and dirty. Going forward, it is also possible to perform multi-classification using vector machines (SVMs) instead of binary classification to use a profile that includes all four attack types. Additionally, exploring other unsupervised search methods beyond DBSCAN may provide insights into improving the accuracy of access searches and other performance metrics. It forms the basis for future research to improve mesh network search efficiency, accuracy, and security capabilities.

References

1. S. Meraji and C. Tropper, "A Machine Learning Approach for Optimizing," no. Icosec, pp. 825– 830, 2010.
2. Al-Khasawneh, M. A., Raza, A., Khan, S. U. R., & Khan, Z. (2024). Stock Market Trend Prediction Using Deep Learning Approach. Computational Economics, 1-32.
3. A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," Futur. Gener. Comput. Syst., vol. 82, pp. 761–768, 2018, doi:
4. 10.1016/j.future.2017.08.043.
5. L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 446–452, 2015, doi: 10.17148/IJARCCE.2015.4696.
6. Khan, U. S., Ishfaque, M., Khan, S. U. R., Xu, F., Chen, L., & Lei, Y. (2024). Comparative analysis of twelve transfer learning models for the prediction and crack detection in concrete dams, based on borehole images. Frontiers of Structural and Civil Engineering, 1-17.
7. V. Kumar, A. Jain, and P. N. Barwal, "Wireless Sensor Networks : Security Issues , Challenges and," vol. 4, no. 8, pp. 859–868, 2014.
8. J. Prajapati and S. C. Jain, "Machine Learning Techniques and Challenges in Wireless Sensor Networks," Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 233– 238, 2018, doi: 10.1109/ICICCT.2018.8473187.
9. Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U., & Liu, J. (2024). Enhancing ASD classification through hybrid attention-based learning of facial features. Signal, Image and Video Processing, 1-14.
10. A. Y. Barnawi and I. M. Keshta, "Energy Management of Wireless Sensor Networks based on Multi-Layer Perceptrons," pp. 939–944, 2023.
11. S. M. Al-tabbakh, "Energy Management of Wireless Sensor Network Based on Modeling by Game Theory Approaches," 2015, doi: 10.1109/EMS.2015.62.
12. S. Lee et al., "An Energy-Efficient Distributed Unequal Clustering Protocol for Wireless Sensor Networks," pp. 443–447, 2008.
13. G. Oddi, A. Pietrabissa, and F. Liberati, "Energy balancing in multi-hop Wireless Sensor Networks : an approach based on reinforcement learning," pp. 262–269, 2014.
14. Y. Sun and L. Li, "Hybrid Learning Algorithm for Effective Coverage in Wireless Sensor Networks," pp. 227–231, 2008, doi: 10.1109/ICNC.2008.320.
15. A. Attiah, M. F. Amjad, M. Chatterjee, and C. C. Zou, "An Evolutionary Game for Efficient Routing in Wireless Sensor Networks," 2016.
16. Khan, S.U.R.; Asif, S.; Bilal, O.; Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. Int. J. Imaging Syst. Technol. 2024, 34, e23044.
17. Raza, A.; Meeran, M.T.; Bilhaj, U. Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers. VFAST Trans. Softw. Eng. 2023, 11, 80–92.
18. Z. Abolfazli, "A Homogeneous Wireless Sensor Network Routing Algorithm : An Energy Aware Cluster Based Approach," no. Icee, pp. 1717–1722, 2014.
19. L. Yang, "A Game Theoretic Approach for Balancing Energy," 2017, doi: 10.3390/s17112654.
20. A. K. Pathan and H. Lee, "Security in Wireless Sensor Networks : Issues and Challenges," pp. 1043–1048, 2006.
21. Khan, U. S., & Khan, S. U. R. (2024). Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT. Multimedia Tools and Applications, 1-21.
22. Khan, S. U. R., & Asif, S. (2024). Oral cancer detection using feature-level fusion and novel self-attention mechanisms. Biomedical Signal Processing and Control, 95, 106437.
23. Dai, Q., Ishfaque, M., Khan, S. U. R., Luo, Y. L., Lei, Y., Zhang, B., & Zhou, W. (2024). Image classification for sub-surface crack identification in concrete dam based on borehole CCTV images using deep dense hybrid model. Stochastic Environmental Research and Risk Assessment, 1-18.
24. Khan, M. A., Khan, S. U. R., Haider, S. Z. Q., Khan, S. A., & Bilal, O. (2024). Evolving knowledge representation learning with the dynamic asymmetric embedding model. Evolving Systems, 1-16.
25. D. G. Padmavathi and M. D. Shanmugapriya, "A Survey of Attacks, Security Mechanisms and Challenges in Wireless Sensor Networks," vol. 4, no. 1, pp. 1–9, 2009, [Online]. Available: http://arxiv.org/abs/0909.0576

26. R. Kozik, M. Choraś, M. Ficco, and F. Palmieri, "A scalable distributed machine learning approach for attack detection in edge computing environments," J. Parallel Distrib. Comput., vol. 119, pp. 18–26, 2018, doi: 10.1016/j.jpdc.2018.03.006.
27. S. Prabavathy, K. Sundarakantham, and S. M. Shalinie, "Design of Cognitive Fog Computing for Intrusion Detection in Internet of Things," vol. 20, no. 3, pp. 291–298, 2018.
28. S. Rathore and J. H. Park, "Semi-supervised learning based distributed attack detection framework for IoT," Appl. Soft Comput. J., vol. 72, pp. 79–89, 2018, doi: 10.1016/j.asoc.2018.05.049.
29. K. R. Choo, "A Two-Layer Dimension Reduction and Two-Tier Classification Model for AnomalyBased Intrusion Detection in IoT Backbone Networks," vol. 7, no. 2, 2019.
30. L. M. Ibrahim, D. B. Taha, and M. S. Mahmod, "(KDD99 , NSL-KDD) BASED ON SELF ORGANIZATION MAP (SOM) ARTIFICIAL NEURAL NETWORK," vol. 8, no. 1, pp. 107–119, 2013.
31. Farooq, M. U., Khan, S. U. R., & Beg, M. O. (2019, November). Melta: A method level energy estimation technique for android development. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE
32. S. U. Jan and S. Ahmed, "Toward a Lightweight Intrusion Detection System for the Internet of Things," IEEE Access, vol. 7, pp. 42450–42471, 2019, doi: 10.1109/ACCESS.2019.2907965.
33. Farooq, M.U.; Beg, M.O. Bigdata analysis of stack overflow for energy consumption of android framework. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–9.
34. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. Int. J. Imaging Syst. Technol. 2024, 34, e22975.
35. Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X.; Zhu, Y. GLNET: Global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides. J. Supercomput. 2023, 80, 7316–7348.
36. J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Semi-Supervised Network Traffic Classification," pp. 369–370, doi: 10.1145/1254882.1254934.
37. Mahmood, F., Abbas, K., Raza, A., Khan, M. A., & Khan, P. W. (2019). Three dimensional agricultural land modeling using unmanned aerial system (UAS). International Journal of Advanced Computer Science and Applications, 10(1).
38. Raza, A., & Meeran, M. T. (2019). Routine of encryption in cognitive radio network. Mehran University Research Journal of Engineering & Technology, 38(3), 609-618.
39. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," 2009 IEEE Symp. Comput. Intell. Secur. Def. Appl., no. Cisda, pp. 1–6, 2009, doi: 10.1109/CISDA.2009.5356528.
40. S. Choi and H. Chae, "Feature Selection using Attribute Ratio in NSL-KDD data," pp. 8–10, 2014. [33] A. Saied, R. E. Overill, and T. Radzik, "Neurocomputing Detection of known and unknown DDoS attacks using Arti fi cial Neural Networks," Neurocomputing, vol. 172, pp. 385–393, 2016, doi: 10.1016/j.neucom.2015.04.101.
41. Khan, S.U.R.; Raza, A.;Waqas, M.; Zia, M.A.R. Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding. Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol. 2023, 7, 10–23.
42. Wajid, M., Abid, M. K., Raza, A. A., Haroon, M., & Mudasar, A. Q. (2024). Flood Prediction System Using IOT & Artificial Neural Network. VFAST Transactions on Software Engineering, 12(1), 210-224.