

A Comprehensive Review on Intrusion Detection System Using ML

Manasa U¹ and Dr. S. Kuzhalivaimozhi²

¹⁻²National Institute of Engineering/ Department of Information Science and Engineering, Mysore, India
Email: 2021scn_manasau_a@nie.ac.in, kuzhali_mozhi@nie.ac.in

Abstract— One of the main issues of the modern day is network security. The vulnerabilities of network security have grown in importance as a result of the internet's recent rapid expansion and widespread use. Unauthorized access and unexpected attacks on secured networks are found using intrusion detection systems. Many studies on the intrusion detection system have been undertaken in recent years. One of the most common and difficult issues that cybersecurity firms confront is intrusion detection. Identification of network security threats is crucially dependent on intrusion detection systems (IDS). For many intranet/internet applications, it safeguards the network from weak source code, viruses, worms, and illegal intrusions. There are still many network security issues even if there are numerous open-source APIs and tools for intrusion detection. When applying self-learning-based classification algorithms, these issues are resolved by performing the correct pre-processing, normalization, feature selection, and ranking on benchmark dataset attributes. This study provides a thorough analysis of the intrusion detection system (IDS). Prior research has examined the IDS based on the methods (algorithms) employed or the categories of intrusions itself.

Index Terms— Intrusion detection, Machine learning, Network security

I. INTRODUCTION

Due to the procedures involved in the electronic transformation of data, the dependence on computer systems and the internet in modern life has resulted in major security, privacy, and confidentiality challenges [1]. The security and privacy of computer systems have undergone significant development, but these issues still persist; in fact, there isn't a system in existence that is 100 percent safe. the I's structure There are several kinds of network assaults as well [2], and these attacks change whenever a new signature with an odd behavior appears in the database of signatures. Many tools are being developed and employed in various sorts of network attacks as a result of the introduction of several attack types. An instrument frequently used to keep track of a network's security is the intrusion detection system (IDS). A variety of network systems, cloud computing platforms, and information systems can all be monitored with this tool. The confidentiality, availability, and integrity of a system's security features can be threatened by assaults, which the IDS can watch for and identify. The purpose of this study is to categorize IDS according to their intended purpose and to compare various IDS types within each category [3].

A. Intrusion Detection System

A software program called the Intrusion Detection System uses different machine learning algorithms to find network intrusions [4,5]. IDS keeps an eye out for malicious behavior and guards against unwanted access from

and "bad connections" (intrusion/attacks) is the goal of the intrusion detector learning challenge. Network traffic is monitored for suspicious activities by an intrusion detection system (IDS), which then warns the system or network administrator. The IDS may occasionally act in response to suspicious or malicious activity, such as preventing the user or source IP address from gaining access to the network [6]. IDS exist in a variety and have various methods for spotting abnormal traffic. Intrusion detection systems can be host-based (HIDS) or network-based (NIDS). Like how antivirus software normally detects and guards against malware, some IDS detect by looking for specific signatures of known threats, while others detect by comparing traffic patterns to a baseline and searching for anomalies. IDS can either monitor and inform users or they can take action in response to a threat they have spotted [7]. The general design of an intrusion detection system is shown in Fig 1.1. Table 1 highlighted machine learning (ML) algorithms and techniques can be further classified into two groups: classical learning and deep learning. Neural Network (NN)-based approaches have gained ground as a result of the recent growth of ML-based solutions. DNN-based techniques typically demand large amounts of computing, however because to recent technological developments, this is not a huge problem [8].

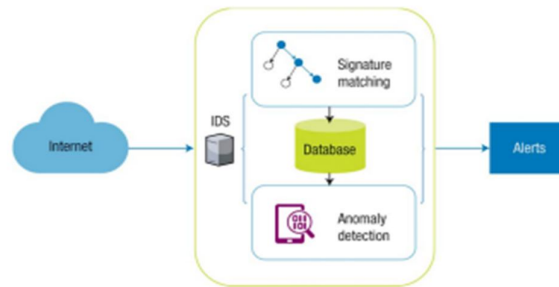


Fig 1.1: Intrusion Detection System Architecture

TABLE I: SUMMARY OF ML ALGORITHM TYPES

Learning Type	Description	Algorithms
Supervised	With supervised learning, the desired outcome is already known. The algorithm is given the labelled dataset so it can learn from it and build a model based on it. Classification and regression-based algorithms are two types of supervised algorithms.	Decision Trees, K-Nearest Neighbour, Linear Regression, Support Vector Machine, Neural Networks
Unsupervised	Unsupervised learning makes use of datasets that lack labels. In order to recognize each data type, the algorithm learns by itself and unearths hidden patterns. The clustering algorithm outlier identification, association rule mining, and auto encoders are further categories under which an unsupervised method might be categorized.	Prior, K-means clustering
Reinforced	The feedback techniques and experiences are how reinforced algorithms learn. It may be argued that learning happens through trial and error.	Q-Learning, Deep Adversary Networks, Temporal Difference
Semi-Supervised	Algorithms for semi supervised learning can learn from both labelled and unlabeled input. When labelling data is challenging or when it is necessary to extract important features, this strategy is applied (i.e., CT scan, MRIs where experts can identify some points to identify a medical condition but may not be able to identify every indicator).	Generative models, Low-density separation, Graph-based methods, Heuristic approaches

B. Classification of Intrusion Detection System

Network Intrusion Detection System (NIDS):

Network intrusion detection systems (NIDS) are installed at a predetermined location within the network to monitor all network traffic [17]. It carries out an observation of all subnet traffic passing through and compares that traffic to a database of known attacks. The alert can be delivered to the administrator as soon as an attack is detected or unusual behavior is noticed.

Host Intrusion Detection System (HIDS):

Host intrusion detection systems (HIDS) are network applications that run on separate hosts or gadgets. Only the incoming and outgoing packets from the device are monitored by a HIDS, which notifies the administrator of

any unusual or malicious behavior. It takes a snapshot of the system files that are currently in use and compares it to the prior snapshot [18]. An alert is given to the administrator to look into if the analytical system files were altered or deleted.

Protocol-based Intrusion Detection System (PIDS):

A system or agent that consistently resides at the front end of a server, regulating and interpreting the protocol between a user/device and the server, makes up a protocol-based intrusion detection system (PIDS). By continuously monitoring the HTTPS protocol stream and accepting the associated HTTP protocol, it tries to secure the web server. As HTTPS isn't encrypted and doesn't immediately enter the web presentation layer, the system would need to be located within this interface in order to use HTTPS.

Application Protocol-based Intrusion Detection System (APIDS):

A system or agent called Application Protocol-based Intrusion Detection System (APIDS) typically resides within a server cluster. By observing and analyzing communication on application-specific protocols, it detects intrusions. For instance, this would keep track of the SQL protocol that the middleware explicitly uses when communicating with the web server's database.

Hybrid Intrusion Detection System:

A hybrid intrusion detection system is created by combining two or more intrusion detection system methodologies. Host agent or system data is merged with network data in the hybrid intrusion detection system to create a comprehensive picture of the network system. In comparison to other intrusion detection systems, hybrid systems are more effective. Hybrid IDS is demonstrated by Prelude.

C. Detection Method of IDS:

Signature-based Method:

The number of bytes, ones or zeros in the network traffic are only a few examples of the specific patterns that signature-based IDS uses to identify attacks. It identifies based on the malware's already-known harmful instruction sequence. Signatures are the patterns that the IDS has identified [6]. IDS that use signatures can quickly identify attacks whose signatures are already known to the system, but it might be challenging to identify newly discovered malware attacks whose signatures are unknown.

Anomaly-based Method:

While new malware is generated quickly, anomaly-based IDS was introduced to identify unknown malware attacks [17]. In anomaly-based IDS, machine learning is used to build a reliable activity model that is compared to anything arriving and is labeled suspicious if it is not found in the model. While these models may be trained based on the applications and hardware setups, machine learning-based approaches offer a greater generic property than IDS that uses signatures.

D. Intrusion Prevention System

The terms intrusion detection and prevention system and intrusion prevention system are interchangeable. It is a network security tool that keeps an eye out for suspicious behavior on systems or networks [20]. The main duties of intrusion prevention systems are to spot harmful activity, gather data on it, report it, and make an effort to block or stop it. Because both IPS and IDS monitor network traffic and system operations for malicious behavior, intrusion prevention systems are thought of as an addition to intrusion detection systems (IDS).

II. LITERATURE SURVEY

By utilizing supervised, semi-supervised, and unsupervised learning algorithms [1] like support vector machines, k-nearest neighbors, neural networks, deep neural networks, recurrent neural networks, convolutional neural networks, deep belief networks, restricted Boltzmann machines, deep Boltzmann machines, and deep auto encoders, machine and deep learning approaches play a crucial role in the detail analysis of anomaly-based intrusion detection. Several researchers have used these machine learning algorithms to analyze the best performance indicator metrics, including accuracy, precision, recall, and positive/false alert rate, on publicly accessible benchmark datasets. The following are some noteworthy earlier studies.

In paper [2], the authors describe a system for classifying cyberattacks that makes use of a deep neural network model. Three phases make up the system: 1) Data gathering 2) Pre-processing of data. 3) classification using deep neural networks. With a learning rate of 0.01, training epochs 10, and input units 86, they exploited global

optical parameters to obtain high performance accuracy of approximately 96.30%. The outcomes demonstrate that SVM model outperforms other conventional machine learning techniques, such as k-nearest neighborhood, linear regression, and random forest. Investigated the anomaly-based intrusion detection system [2,3]. Using the NSL- KDD dataset, they created a model for anomaly-based intrusion detection using a variety of machine and deep learning algorithms. They tested deep convolutional neural network (DCNN) and LSTM models against the traditional machine learning classification techniques k closest neighbor, SVM, Random Forest, and Decision Tree, and reported approximately 85-89% accuracy on the NSL-KDD test dataset.

They evaluated how the NSL-KDD dataset performed using a variety of machine learning classification techniques, such as Naive-Bayes, Support Vector Machines, and Decision-Trees. Around 91% accuracy was recorded by the hybrid detector [4]. For intrusion detection on the NSL-KDD dataset, authors presented a model that combines the restricted Boltzmann machine and support vector machine classifiers [5- 7]. In the training set, they chose 22 attack kinds, and in the testing set, they chose 17 attack types. When compared to conventional support vector machines, the model produced the best performance [8].

Before using classification algorithms like SVM, k-NN, NN, DT, and NB, various feature selection approaches and classification techniques are used with CFS, PCA, and IGR to reduce the duplication in data attributes [9]. The experimental results on the NSL-KDD dataset demonstrate that k-NN classifiers outperform other classifiers in terms of performance, and that the Information Gain Ratio (IGR) feature selection method outperforms other feature selection methods in terms of performance [10]. An IDS receives fresh data from sensors, examines this data, and then takes action [11].

The technique of gleaning knowledge from vast amounts of data is known as machine learning. Machine learning models are made up of a collection of guidelines, techniques, or sophisticated "transfer functions" that can be used to recognize or forecast behavior, as well as to uncover interesting data patterns [12].

The field of AIDS has seen substantial use of machine learning techniques. For extracting knowledge from intrusiondatasets, a variety of algorithms and approaches have been used, including clustering, neural networks, association rules, decision trees, genetic algorithms, and closest neighbor methods [13]. To stop recent attacks, current detection and prevention techniques need continual upkeep. For security specialists, maintaining the aforementioned systemsalso entails updating and reviewing logs ona daily basis. Because of the technological advancements in the online world, attacker techniques have also altered, making the traditional tools more susceptible. Bypassing and evading security measures ismade possible by such liabilities, which enable intruders to carry out harmful operations [14]. The authors covered a range of supervised and unsupervised learning approaches for machine learning and deep learning. The authors of the publications [15] have outlined a methodology in which severaldata sources, including logs, packet flows, and sessions, are gathered and provided to ML algorithms. However, a suitable standardized data collection must be available in order to attain high accuracy.

Machine learning, deep learning, and data mining are just a few of the methods that have been examined by many authors and applied in NIDS; these methods are known as anomaly-based NIDS. However, the most recent study has proposed a hybridstrategy that uses NIDS that are based on both anomaly and signature data. Many types of network assaults, including DoS, DDoS, probe attacks, IP spoofing, etc., are protected using these strategies. The detection rate of network attacks is another area where deep learning performs well, and this rate can be raised by combining it with a hybrid approach and employing machine learning classifiers [16].

TABLE II: COMPARISON OF RESULTS

Sl.No	Technique	Dataset	Accuracy(%)
1	NB andNaive Bayes	4,266 programs;3,265 malware and1,001 benign programs	97.76
2	GBM	Microsoft malware classificationchallenge dataset	96
3	KNN, GB, NB, andDT	21,740 malwares	98.9
4	NB, SVM	29,580 files	93.07
5	Logistic Regression	Creates byauthors	98
6	KNN, NB,SVM	250 Indonesian malware	96.8
7	NB, IBK	63,000 files	99.97
8	Decision tree andRandom Forest	Malware from OpenMalware	96
9	Deep Belief Networks	192 malicious apps	96.76
10	SVM	2000 android applications, 1000 malware	95.2

III. METHODOLOGY

The reference papers have introduced many algorithms for our detection, this section gives a description about the algorithms or methods used and their related accuracy also authors performed different kind of classification such as multi-class rather than binary methods. Table 2 and Table 3 shows the accuracy comparison between various techniques with respect to parameters obtained for Intrusion detection

TABLE III: COMPARISON OF ALGORITHMS BASED ON VARIOUS PARAMETERS

Authors	Algorithm	Methodology	Advantage	Disadvantage
Yin et al.[22]	Deep Learning	Recurrent Neural Network	A NIDS based on RNN model is proposed that achieves higher performance in terms of detection accuracy using ML algorithms	The model has increased complexity and requires more time for training. Also, it uses an older dataset NSL-KDD for model evaluation.
Gu et al. [23]	Machine Learning	K-Means	True positive rate in detection is high	False positive rate is unstable
Chen et al. [24]		SVM, NB and DT	Perform well with highly imbalance training data	Performance is unstable when the imbalance data ratio is 1000
Naseer et al. [28]	Machine Learning and Deep Learning	Comparison between different ML and DL-based IDS Models	Comparison of ML- and DL-based NIDS algorithms by implementing on GPU-integrated testbed.	Use of an older dataset NSL-KDD for evaluation.
Ravi et al.[29]	Machine Learning	NN and K-Means	Higher accuracy than some other models	Data set used in validation is too old
Jia et al.[27]	Deep Learning	Deep Neural Network	Proposed a NIDS based on DNN with four hidden layers that have shown superior performance than IDS based on ML methods.	For validating the model, older datasets KDD cup'99 and NSLKDD are used. Also, lower detection accuracy for the U2R attack class is recorded.
Yu et al. [26]	Deep Learning	Few Shots Learning (FSL) using DNN and CNN	An efficient NIDS model is proposed based on the novel DL idea of Few-shot Learning (FSL). DL algorithms as DNN and CNN are adopted as embedding functions in the model for extracting the essential feature and reducing the dimension.	For the NSL-KDD dataset, the model achieves reasonable detection rates for U2R and R2L attacks but it is still lower comparing the other attack classes of the dataset.
Marir et al. [25]	Machine Learning and Deep Learning	Deep Belief Network and SVM	Use of DBN and ensemble SVM to detect abnormal behavior in a distributed way. DBN is used for feature extraction, which is then forwarded to the ensemble SVM and finally, the voting mechanism is used for prediction.	Use of DBN and ensemble SVM to detect abnormal behavior in a distributed way. DBN is used for feature extraction, which is then forwarded to the ensemble SVM and finally, the voting mechanism is used for prediction.

Following is the comparison of these algorithms referring to various papers

Algorithm	
Random Forest	Precision – 90.04%, Recall- 82.08%, Accuracy –84.62%
Cluster Center andNearest Neighbor	Precision – 93.38%, Recall- 79.9%, Accuracy- 88.05%
Decision Tree and SVM	Precision- 86.82%, Recall- 90.64%, Accuracy- 86.84%
Logistic Regression	Precision -95.11%, Recall -82.96%, Accuracy-83.06%
Neural Networks	Precision- 94.15%Recall-78.28%,Accuracy- 80.98%
GradientBoosted Decision Tree	Precision-96.31% Recall-95.55% Accuracy-94.83%

IV. CONCLUSION

The need to address computer network security is becoming more and more important in today's society. Conventional intrusion detection methods are not flexible enough to adapt to changing networks or to unanticipated assault types. Techniques for machine learning have been shown to be effective at detecting intrusions. Machine learning methods paired with feature engineering and other elements can result in high accuracy in intrusion detection. Some of them include choosing the suitable feature set, choosing the right training and testing data, etc. A higher performance could be attained by choosing the proper qualities for these elements. A multilevel hybrid classification intrusion detection system is suggested in this research paper. Comparing the suggested approach to single level classification models, it may perform better.

REFERENCES

- [1] Lee, Sang-Woong, et al. "Towardssecure intrusion detection systems using deep learning techniques: Comprehensive analysis and review." *Journal of Network and Computer Applications* 187 (2021): 103111.
- [2] Lansky, Jan, et al. "Deep learning-based intrusion detection systems: a systematic review." *IEEE Access* 9 (2021): 101574-101599.
- [3] Salih, Azar Abid, et al. "Deep learning approaches for intrusion detection." *Asian Journal of Research in Computer Science* (2021): 50-64.
- [4] Liu, Hongyu, and Bo Lang. "Machine learning and deep learning methods for intrusion detection systems: A survey." *applied sciences* 9.20 (2019): 4396.
- [5] Ahmad, Zeeshan, et al. "Network intrusion detection system: A systematic study of machine learning and deep learning approaches." *Transactions onEmerging TelecommunicationsTechnologies* 32.1 (2021): e4150.
- [6] B. Liu H, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey,," MDPI, Applied Sciences, vol. 9, no. 20, p. 4396, 2019.
- [7] STH, "Types of Machine Learning: Supervised Vs Unsupervised Learning,," Software Testing Help, 30 June 2020. [Online]. Available: <https://www.softwaretestinghelp.com/type-s-of-machine-learning-supervised-unsupervised/>. [Accessed 22 July 2020].
- [8] P. M. & R. J. Drewek-Ossowicka, "A survey of neural networks usage for intrusion detection systems," *Journal of Ambient Intelligence and Humanized Computing* (2020), 2020.
- [9] Kocher, Geeta, and Gulshan Kumar. "Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges." *Soft Computing* 25.15 (2021): 9731-9763.
- [10] Vanin, Patrick, et al. "A study of network intrusion detection systems using artificial intelligence/machine learning." *Applied Sciences* 12.22 (2022): 11752.
- [11] Sharma, Rupam Kr, Hemanta Kumar Kalita, and Parashjyoti Borah. "Analysis of machine learning techniques-based intrusion detection systems." *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics: ICACNI 2015, Volume 2*. Springer India, 2016.
- [12] Nassar, Mostafa, et al. "Network intrusion detection, literature review and some techniques comparison." *2019 15th International Computer EngineeringConference (ICENCO)*. IEEE, 2019.
- [13] Rekha, Gillala, et al. "Intrusion detection in cyber security: role of machine learning and data mining in cyber security." *Advances in Science, Technology and Engineering Systems Journal* 5.3 (2020): 72-81.
- [14] Rupa Devi, T., and Srinivasu Badugu. "A review on network intrusion detection system using machine learning." *Advances in Decision Sciences, Image Processing, Security and Computer Vision: International Conference on Emerging Trends in Engineering (ICETE)*, Vol. 2. Springer International Publishing, 2020.
- [15] Alshamy, Reem, and Mossa Ghurab. "A review of big data in network intrusion detection system: Challenges, approaches,datasets, and tools." *Journal of Computer Sciences and Engineering* 8.7 (2020):62- 74.
- [16] Ahmad, Zeeshan, et al. "Network intrusion detection system: A systematic study of machine learning and deep learning approaches." *Transactions onEmergingTelecommunicationsTechnologies* 32.1 (2021): e4150.
- [17] Lakshminarayana, Deepthi Hassan. *Intrusion detection using machine learning algorithms*. East Carolina University, 2019.
- [18] Yang, Li, and Abdallah Shami. "IDS- ML: An open-source code for Intrusion Detection System development using Machine Learning." *Software Impacts* 14 (2022): 100446.

- [19] A Meshal, R. A. K. "A comparative study for Intrusion Detection Metods Using Machine Learning." 13.
- [20] Kathiresan, V., et al. "A Comparative Study of Diverse Intrusion Detection Methods using Machine Learning Techniques." 2022 InternationalConference on Computer Communicationand Informatics (ICCCI). IEEE, 2022.
- [21] <https://www.semanticscholar.org/paper/r/A-Comparative-Analysis-of-Machine-Learning-for-and-Al-Janabi-Altamimi/8341e7f8527b113bedfc51e6b328d62b74f480fc>
- [22] Yin C, Zhu Y, Fei J, He X. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*.2017;5:21954-21961.<https://doi.org/10.1109/ACCESS.2017.2762418>.
- [23] Y. Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised K-means DDoS detection method using hybrid feature selection algorithm," *IEEE Access*, vol. 7, pp. 64351–64365, 2019
- [24] Z. Chen, Q. Yan, H. Han, S. Wang, L. Peng, L. Wang, and B. Yang, "Machine learning based mobile malware detection using highly imbalanced network traffic," *Inf. Sci.*, vols. 433–434, pp. 346–364, Apr. 2018.
- [25] Marir N, Wang H, Feng G, Li B, Jia M.Distributed abnormal behavior detectionapproach based on deep belief network andensemble svm using spark. *IEEE Access*.2018;6:59657-59671. <https://doi.org/10.1109/ACCESS.2018.2875045>.
- [26] Yu Y, Bian N. An intrusion detection method using few-shot learning. *IEEE Access*.2020;8:49730-49740. <https://doi.org/10.1109/ACCESS.2020.2980136>.
- [27] Jia Y, Wang M, Wang Y. Networkintrusion detection algorithm based on deepneural network. *IET Inf Secur*.2018;13(1):48-53. <https://doi.org/10.1049/iet-ifs.2018.5258>.
- [28] Naseer S, Saleem Y, Khalid S, et al. Enhanced network anomaly detection based on deep neural networks. *IEEE Access*.2018;6:48231-48246.<https://doi.org/10.1109/ACCESS.2018.2863036>.
- [29] N. Ravi and S. M. Shalinie, "Semisupervised-learning-based security to detect and mitigate intrusions in IoT network," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 11041–11052, Nov. 2020.