ANMACHINE LEARNING PROCEDURE FOR DISTRIBUTING SECURITY SAVING AREA INFORMATION

Madduri Deepika[1]., Chanda Sathwika[2].,Nalumachu Sanjana[3].,Vanavasam Jaya Sree[4].,Priyanka Pandey[5]

1 Assistant Professor, Department Of CSE., Malla Reddy College Of Engineering For Women., Maisammaguda.,

Medchal.,Ts, India (✉madduridеepika9644@gmail.com)

2, 3, 4, 5 B.Tech CSE, (19RG1A05J4, 19RG1A05M5, 19RG1A05P7, 19RG1A05N3),

Malla Reddy College Of Engineering For Women., Maisammaguda., Medchal., Ts, India

Abstract:

Cyber security has emerged as a crucial topic of study due to the pervasiveness of networks in contemporary society. An intrusion detection system (IDS) is a critical component of network security since it keeps tabs on all the programs and devices connected to the internet. Existing IDSs have come a long way over the years, but they still have some work to do when it comes to identifying novel threats, decreasing false positives, and boosting detection accuracy. To address these issues, several studies have centred on the creation of IDSs that take use of machine learning techniques. Automatically and with a high degree of precision, machine learning techniques can identify the key distinctions between "normal" and "abnormal" data. Machine learning techniques may also identify previously unseen assaults because of their high generalizability. Deep learning is a subfield of machine learning that has attracted a lot of attention due to its impressive results. In order to identify and describe the machine learning-based and deep learning-based IDS literature, a taxonomy is proposed in this review. We think this taxonomy structure is ideal for anybody studying cybersecurity. The survey begins with a definition and classification of IDSs. Then, we dive into the common machine learning algorithms found in intrusion detection systems, metrics, and standard datasets for comparison. After establishing the suggested taxonomy as a foundation, we combine it with sample literature to detail how to use machine learning and deep learning to address significant challenges in IDS. Finally, we evaluate recent representative studies to identify difficulties and potential advancements.

Keywords:

Keywords: IDS, Machine Learning, Deep Learning, Cyber Security

Introduction

Cyber security is becoming more crucial as the role of networks expands in contemporary society. Anti-virus programs, firewalls, and intrusion detection systems (IDSs) are the mainstays of cyber security measures. These methods prevent malicious actors from gaining access to networks. Among them is the intrusion detection system (IDS), a sort of detection system that is crucial to preserving network security by keeping tabs on the configurations of all the devices and programs connected to it. In 1980 [1], the concept of an intrusion detection system was originally presented. Many established IDS products have emerged since then. However, many IDSs still have a high false alarm rate, producing many alerts for low nonthreatening situations, which increases the burden on security analysts and may lead to the disregard of attacks that are seriously harmful. As a result, a lot of effort has gone into improving IDSs with better detection and fewer false positives. Existing IDSs also have the issue of not being able to identify assaults they have never seen before. Due to the dynamic nature of networks, new forms of attack and variations on existing ones are always appearing. IDSs that can identify previously unseen

assaults must thus be developed. Researchers have started to concentrate on building IDSs using machine learning techniques in an effort to solve the aforementioned issues. Machine learning is an AI method that can automatically mine large data sets for relevant insights [2]. When enough training data is made available, machine learning-based IDSs may reach reasonable detection levels, and machine learning models have enough generalizability to identify attack variations and new assaults. In addition, IDSs based on machine learning need nothing in the way of specialized skills to develop and deploy. When it comes to machine learning, deep learning is the most promising area. Deep learning methods are superior to more conventional machine learning approaches when it comes to processing massive datasets. Also, deep learning techniques are efficient and effective because they can automatically learn feature representations from raw data and then output results. The deep structure of many hidden layers is a defining feature of deep learning. Support vector machines (SVMs) and k-nearest neighbour (KNNs) are two examples of older machine learning models that include no or just one hidden layer. Traditional machine learning models are sometimes known as shallow models for this reason.

Identifying and Classifying IDS

For an intrusion detection system, an intrusion is any effort to illegally obtain data from a computer system or to disrupt its normal functioning. An intrusion detection system (IDS) is a program

designed to identify security breaches, both external (such as attempted break-ins) and internal (such as system infiltration and misuse) [6]. IDSs are primarily used for monitoring hosts and networks, analysing computer system behaviour, alerting on suspicious activity, and taking corrective action. Network nodes (such as switches in key network segments) are frequently placed near IDSs since IDSs monitor connected hosts and networks. Both detection-based and data-source-based approaches may be used to categorize IDS systems. IDSs may be classified based on their detection techniques into two categories: misuse detection and anomaly detection. Host-based and network-based techniques are two categories of data-source-based approaches to IDSs [7]. This study combines these two categories of IDS classification approaches, prioritizing the data source and giving less weight to the detection technique. Illustration of the suggested taxonomy. The survey focuses on machine learning techniques for detection. In Section 4, we present the basics of using machine learning to IDS with various data sources.

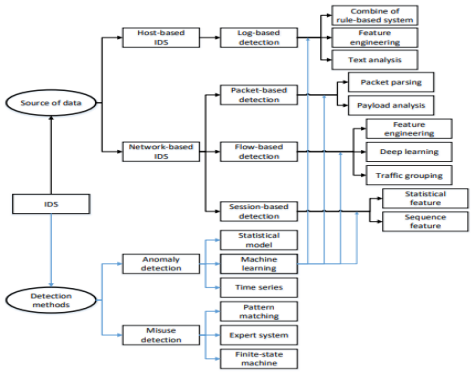

Figure 1. Taxonomy system of IDS.

Classification by Detection Methods

Signature-based detection is another name for misuse detection. The fundamental concept of using signatures to symbolize malicious actions. The samples' signatures are compared to a signature database in the detection procedure. The development of reliable signatures is the primary challenge in creating systems for abuse detection. Misuse detection has a low false alarm rate and provides detailed details on attack types and potential causes; but it also has a high missed alert rate, cannot identify new assaults, and necessitates the maintenance of a large signature database. Anomaly identification is based on the principle of defining anomalous behaviour in terms of how far they deviate from a predetermined standard. Therefore, defining a normal profile explicitly is the first step in developing an anomaly detection system. Anomaly detection has a high false alarm rate and can't identify probable causes for an irregularity, but its strengths include good generalizability and the capacity to spot new threats. Comparison between Misuse Detection with Anomaly Detection Table 1: Key Differences

Table 1. Differences between misuse detection and anomaly detection.

	Misuse Detection	Anomaly Detection
Detection performance	Low false alarm rate; High missed alarm rate	Low missed alarm rate; High false alarm rate
Detection efficiency	High, decrease with scale of signature database	Dependent on model complexity
Dependence on domain knowledge	Almost all detections depend on domain knowledge	Low, only the feature design depends on domain knowledge
Interpretation	Design based on domain knowledge, strong interpretative ability	Outputs only detection results, weak interpretative ability
Unknown attack detection	Only detects known attacks	Detects known and unknown attacks

As shown in Figure 1, in detection method-based taxonomy, misuse detection includes pattern matching-based, expert system, and finite state machine-based methods. Anomaly detection includes statistical model-based, machine learning-based, and time series-based methods.

Data Source Categorization

Since host-based IDSs are able to keep tabs on the doings of important objects, like sensitive files, programs, and ports, they are better able to detect intrusions and take corrective action. Host-based intrusion detection systems can't detect network threats, use up host resources, and rely on the stability of the host. The majority of the time, a network-based IDS is installed on pivotal hosts or routers. The vast majority of network-based IDSs do not rely on any particular OS, making them universally applicable. Also, network-based IDSs can identify certain varieties of protocol and network attacks. One problem is that they can only keep tabs on the data traveling across a limited portion of a network. Table 2 summarizes the key distinctions between host-based and network-based intrusion detection systems.

Table 2. Differences between host-based and network-based IDSs.

	Host-Based IDS	Network-Based IDS
Source of data	Logs of operating system or application programs	Network traffic
Deployment	Every host; Dependent on operating systems; Difficult to deploy	Key network nodes; Easy to deploy
Detection efficiency	Low, must process numerous logs	High, can detect attacks in real time
Intrusion traceability	Trace the process of intrusion according to system call paths	Trace position and time of intrusion according to IP addresses and timestamps
Limitation	Cannot analyze network behaviors	Monitor only the traffic passing through a specific network segment

As shown in Figure 1, a host-based IDS uses audit logs as a data source. Log detection methods are mainly hybrids based on rule and machine learning, rely on log features, and use text analysis-based methods. A network-based IDS uses network traffic as a data source—typically packets, which are the basic units of network communication. A flow is the set of packets within a time window, which reflects the network environment. A session is a packet sequence combined on the basis of a network information 5-tuple (client IP, client port, server IP, server port, protocol). A session represents high-level semantic information of traffic. Packets contain packet headers and payloads; therefore, packet detection includes parsing-based and payload analysis-based methods. Based on featureextraction, flow detection can be divided into feature engineering-based and deep learning-based methods. In addition, traffic grouping is a unique approach in flow detection. Based on whether sequence information is used, session detection can be divided into statistical feature-based and sequence feature-based methods.

IDS Machine Learning Models Frequently Used Algorithms

Supervised learning and unsupervised learning are the two primary categories of machine learning. Labelled data is crucial for supervised learning. While classification is the most common supervised learning task (and the most common IDS application), manually labelling data is time-consuming and costly. Therefore, the primary obstacle to supervised learning is the dearth of labelled data. In contrast, unsupervised learning makes it considerably simpler to collect training data by gleaning useful feature information from unlabelled data. However, unsupervised learning methods typically have worse detection performance than supervised learning methods. Figure 2 depicts a selection of the most popular machine learning methods used by IDSs.

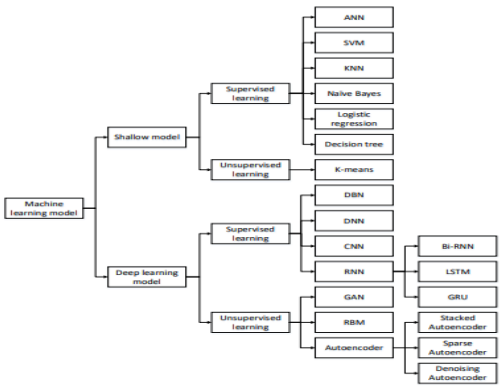

Figure 2. Taxonomy of machine learning algorithms

Detection Systems that Use Machine Learning

Machine learning is an example of a data-driven approach, the first stage of which is to comprehend the data. According to Figure 1, we base our primary categorization scheme on the nature of the data source itself. Here, we propose a number of machine learning-based approaches to IDS design for diverse data kinds. Both host and network-specific attack behaviours are reflected in the various data types. Both host and network actions are recorded in system logs and network traffic, respectively. Each distinct kind of assault has its own distinct pattern. So, in order to detect various attacks based on attack characteristics, it is necessary to select suitable data sources. In order to identify a denial-of-service (DOS) assault, flow data may be used, since this kind of attack is characterized by the rapid transmission of numerous packets. Detection of a covert channel, in which information is leaked between two distinct IP addresses, is easier to do using session data 4.Identifying Attacks in Individual Packets Information about a communication is stored in "packets," the fundamental units of network communication. Since packets are made up of binary data, they are unintelligible until they are parsed. The header and application data make up what is known as a packet. The headers are organized data fields that include protocol-specific information like IP addresses and port numbers. The payload for the protocols operating at the application layer is included in the application data section. Using packets as IDS data sources has three benefits: (1) Packets carry the contents of communications and may be used to successfully identify U2L and R2L attacks because of this. Because (2) packets carry Internet Protocol addresses and timestamps, attackers' origins can be pinpointed with great accuracy. Thirdly, without caching, packets may be analysed instantaneously, allowing for real-time detection. However, it is difficult to detect some attacks, such as DDOS, because individual packets do not reflect the full communication state nor the contextual information of each packet. The most common types of packet-based detection are packet parsing and payload analysis.

Detection Via Packet Parsing

Protocols like Hypertext Transfer Protocol (HTTP) and Domain Name System (DNS) are utilized in network communications. These protocols use various formats, and the fields in the protocol header are the primary focus of detection techniques that rely on packet parsing. In most cases, the header fields are extracted with the help of parsing tools (like Wireshark or the Bro), and the values of the most significant fields are then used

as feature vectors. Methods based on packet parsing may be used for detection in shallow models. To identify assaults, features may be taken from the header data and utilized in classification algorithms. An SVM and K-means based packet identification approach was suggested by Mayhew et al. [40]. Using Bro, they analysed packets taken from a functioning business network. The packets were first sorted by protocol. The K-means++ technique was then used to perform data clustering across all of the available protocol datasets. As a result, the primary dataset was divided into several clusters, with each cluster containing data that was similar to the others in the same way. After that, they trained SVM models on each cluster using the collected characteristics from the packets. In terms of accuracy, they achieved 96% for the HTTP protocol, 93% for TCP, 98% for Wiki, 99% for Twitter, and 93% for email. To combat the issue of a high false alarm rate in packet parsing-based detection, unsupervised learning is often used. A fuzzy C-means based packet identification approach was suggested by Hu et al. [41]. The fuzzy C mean technique modifies the traditional K-means algorithm by using fuzzy logic, changing the way samples are assigned to clusters from a 0/1 Boolean value to a membership degree. In order to extract Snort alerts, source IP addresses, destination IP addresses, source ports, destination ports, and timestamps, they ran the DARPA 2000 dataset through Snort. Next, they clustered the packets based on the information in the feature vectors to determine whether or not they were false alarms. They repeated the clustering algorithms 10 times to cut down on the impact of the first run. The findings indicated that the false alarm rate was decreased by 16.58% and the missed alert rate was decreased by 19.23% thanks to the fuzzy C-means algorithm.

Difficulties and Possible Future Paths

Papers introducing IDSs based on machine learning are included in Table 5 of this overview. The fact that 14 of the 26 publications included here use deep learning techniques demonstrates that this topic is now at the forefront of academic interest. Popular datasets include KDD99 and NSL-KDD. While there has been significant progress in the use of machine learning techniques for intrusion detection, several obstacles remain.

Table 1 Methods and papers on machine learning based IDSs.

Methods	Papers	Data Sources	Machine Learning Algorithms	Datasets
Packet parsing	Mayhew et al. [40]	Packet	SVM and K-means	Private dataset
	Hu et al. [41]	Packet	Fuzzy C-means	DARPA 2000
Payload analysis	Min et al. [43]	Packet	CNN	ISCX 2012
	Zeng et al. [44]	Packet	CNN, LSTM, and autoencoder	ISCX 2012
	Yu et al. [45]	Packet	Autoencoder	CTU-UNB
	Rigak et al. [46]	Packet	GAN	Private dataset
Statistic feature for flow	Goeschel et al. [47]	Flow	SVM, decision tree, and Naïve Bayes	KDD99
	Kuttranont et al. [48]	Flow	KNN	KDD99
	Peng et al. [13]	Flow	K-means	KDD99
Deep learning for flow	Potluri et al. [49]	Flow	CNN	NSL-KDD and UNSW-NB15
	Zhang et al. [50]	Flow	Autoencoder and XGBoost	NSL-KDD
	Zhang et al. [51]	Flow	GAN	KDD99
Traffic grouping	Teng et al. [52]	Flow	SVM	KDD99
	Ma et al. [53]	Flow	DNN	KDD99 and NSL-KDD
Statistic feature for session	Ahmim et al. [54]	Session	Decision tree	CICIDS 2017
	Abelari et al. [55]	Session	K-means	Private dataset
Sequence feature for session	Yuan et al. [56]	Session	CNN and LSTM	ISCX 2012
	Radford et al. [57]	Session	LSTM	ISCX IDS
	Wang et al. [58]	Session	CNN	DARPA 1998 and ISCX 2012
Rule-based	Meng et al. [59]	Log	KNN	Private dataset
	McElwee et al. [60]	Log	DNN	Private dataset
Log feature extraction with sliding window	Tran et al. [61]	Log	CNN	NGIDS-DS and ADFA-LD
	Tuor et al. [62]	Log	DNN and RNN	CERT Insider Threat
	Bohara et al. [63]	Log	K-means and DBSCAN	VAST 2011 Mini Challenge 2
Text analysis	Uwagbole et al. [64]	Log	SVM	Private dataset
	Vartouni et al. [65]	Log	Isolate forest	CSIC 2010 dataset

One) There aren't enough data sets to use. KDD99 is now the most used dataset, although it has several issues and new datasets are needed. However, building new datasets requires specialized knowledge and requires substantial human labour. The insufficiency of datasets is made worse by the unstable nature of the Internet. Some existing datasets are too old to reflect the emergence of new types of attacks. In a perfect world, datasets would reflect the state of the art in network security and contain the vast majority of known threats. In addition, the existing datasets have to be representative, balanced, have less duplication, and have less noise. Possible solutions to this issue include incremental learning and the systematic creation of datasets. Less reliable detection in natural settings. While machine learning techniques can be useful for intrusion detection, they struggle to make accurate predictions when presented with completely novel data. Most previous research has relied on labelled datasets. Even though the models are very accurate on test sets, this does not ensure they will perform well in real-world settings if the dataset does not include all representative samples of the target environment. Poor performance. investigations often involve intricate models and significant data preparation procedures, which leads to inefficient detection findings, as is the case with the vast majority of investigations. However, IDSs need to detect attacks in real time to minimize damage. Accordingly, there is a compromise between effectiveness and efficiency. Common approaches include parallel computing [66,67] and GPUs [48,68,69]. After reviewing the literature, we can identify the following as the primary directions of IDS study. Making use of (1) specialized information. When trying to identify very particular kinds of assaults in very specific kinds of application situations, combining domain knowledge with machine learning might increase the detection result.

• Rule-based detection approaches need a great deal of specialist knowledge, yet they have low false alarm rates and high missed alarm rates. However, machine learning approaches frequently exhibit low missed alarm rates and high false alarm rates. Each approach has benefits that complement the

other. IDSs with low false alarm rates and low missed alarm rates may be achieved by combining machine learning approaches with rule-based systems like Snort [70-73]. The right feature must be extracted for distinct forms of assaults including denial-of-service [74–79], botnet [80], and phishing web [81], where the attack characteristics are abstracted using domain knowledge. Cloud computing [82,83], the Internet of Things [84–86], and smart grids [87,88] are just a few examples of the types of application scenarios where domain knowledge may be leveraged to give the environmental features that are beneficial in data collecting and data preparation. Enhancing learning machine algorithms, 2. Increasing the efficiency of detection relies mostly on developments in machine learning techniques. As a result, there has been a rise in the number of research that use both deep learning and unsupervised learning techniques.

• Deep learning techniques have a greater capacity for fitting data since they learn features directly from the raw data. Classification, feature extraction, feature reduction, data denoising, and data augmentation are just few of the many applications of deep learning models using deep structures. Therefore, IDSs can benefit greatly from deep learning techniques. There is no need for labelled data with unsupervised learning techniques, hence they may be employed even if there is a scarcity of data. Typically, researchers will use an unsupervised learning model to divide their data, label their clusters by hand, and then use supervised learning to train a classification model [89–92]. Three, creating usable examples. High detection accuracy is important, but IDSs also need to be fast and easy to understand. Real-time capability is crucial for threat detection. Therefore, enhancing the effectiveness of machine learning models is a potential area of study. It is also important to minimize the amount of time spent on data collecting and storage. In practice, IDSs benefit much from being interpretable. There is a lot of mystery around machine learning models, particularly deep learning models. These models just report detection success or failure and provide no context for doing so [93]. However, care must be taken when making any choice regarding cyber security. A persuasive output result must have a clear justification. Therefore, it is more practical to use an IDS with high accuracy, high efficiency, and interpretability.

Conclusion

To organize the many machine learning methods in this area, the study first offers an IDS taxonomy that centers on data sources. We use this classification system to examine and explain IDSs and how they are used with different types of data, including logs, packets, flows, and sessions. In order for IDSs to effectively identify attacks, it is crucial to have a data source that is well-suited to doing so. Logs may be used to identify SQL injection, U2R, and R2L attacks because of the rich semantic information they carry. And the contents of communications provided by packets are suitable for identifying U2L and R2L attacks. Flow is an environment model that can identify DOS and Probe attacks throughout the whole network. Sessions, which mirror client-server communication, may be used to identify U2L, R2L, tunnel, and Trojan assaults. The focus of this study is on machine learning methods (particularly deep learning algorithms) and application scenarios for IDSs that use these various data kinds. The use of deep learning models is growing in significance, and research in this area has emerged as a top priority. The performance of IDSs may be enhanced by using numerous deep networks, which are part of deep learning techniques. As opposed to their shallower counterparts, deep learning models excel in both fitting and generalization. In addition, unlike shallow machine learning models, deep learning techniques don't need feature engineering or domain expertise. However, the real-time requirement of IDSs can't always be met because deep learning models take too long to run.

References

[1]. Anderson, J.P. Computer Security Threat Monitoring and Surveillance; Technical Report; James P. Anderson Company: Philadelphia, PA , USA, 1980.

[2]. Michie, D.; Spiegelhalter, D.J.; Taylor, C. Machine Learning, Neurall and Statistical Classification; Ellis Horwood Series in Artificial Intelligence: New York, NY, USA, 1994; Volume 13.

[3]. Buczak, A.L.; Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tutor. 2015, 18, 1153–1176. [CrossRef]

[4]. Xin, Y.; Kong, L.; Liu, Z.; Chen, Y.; Li, Y.; Zhu, H.; Gao, M.; Hou, H.; Wang, C. Machine learning and deep learning methods for cybersecurity. IEEE Access 2018, 6, 35365–35381. [CrossRef]

[5]. Agrawal, S.; Agrawal, J. Survey on anomaly detection using data mining techniques. Procedia Comput. Sci. 2015, 60, 708–713. [CrossRef]

[6]. Denning, D.E. An intrusion-detection model. IEEE Trans. Softw. Eng. 1987, 222–232. [CrossRef]