Intrusion Detection System With Machine Learning Approach: A Survey

Anish Fathima B[1], Pragathi S K[2], Ragavi P[3], Rathi S[4], Reshma K[5]
[1]Assistant Professor, Dept of Electronic and Communication Engineering
[2, 3, 4, 5]Dept of Electronic and Communication Engineering
[1, 2, 3, 4, 5] Sri Krishna college of Engineering and Technology, Coimbatore

Abstract- *In recent decades, information networks' accelerated growth has culminated in a slew of protection concerns, including computer and network intrusions. Intrusion Detection Systems (IDS) are a form of intrusion detection system that IDSs provide tools for identifying and discriminating between disruptive and non-intrusive network packets. Most modern intrusion detection technologies depend heavily on human observers to distinguish between disruptive and non-intrusive network traffic by analyzing server logs or network traffic. Human presence in the identification mechanism has become a non-trivial concern as network traffic data has grown. The system's capacity to operate autonomously over rapidly increasing data in the network is limited by IDS' ability to perform, dependent on human expertise. On the other hand, soft-computing methods will effectively model human expertise and their capacity to interpret the device. Autonomous packet detections are possible thanks to intrusion detection strategies focused on machine learning and soft computing. They can analyze data packets on their own. These methods are heavily focused on mathematical data processing. The algorithms that deal with these datasets will make judgments based on previous data trends to deal with new emerging data patterns in network traffic. This paper provides a thorough survey of numerous soft-computing and machine learning strategies to develop autonomous IDSs.*

Keywords- IDS, Data Mining, Analysis, classification

I. INTRODUCTION

Modern information management networks depend heavily on intrusion detection systems (IDSs) [1]. In summary, an IDS's function is to detect and recognize any harmful behaviours in a computer system, enabling administrators to respond rapidly and efficiently, either manually or automatically [3].

Almost everybody in the world now has access to the internet. Any consumer has at least 1-2 computers such as laptops or cell phones, and the amount of internet users is growing by the day. When the number of users rises, so do the number of attacks on their computers, especially network-based attacks known as "Network Attacks" [8].

Under the IDS categories, misuse identification differs from anomaly detection. It analyses the data it collects and compares it to vast collections of assault signatures to detect misuse. In anomaly detection, it watches network segments to equate their condition to a standard baseline and check for abnormalities. Because of the vast limitations of computer systems and attackers' ingenuity, misuse detection is a particularly difficult issue. When an alert is activated for regular operation, it is considered a false-positive. When there is no warning about an irregular activity, it is called a false-negative [7].

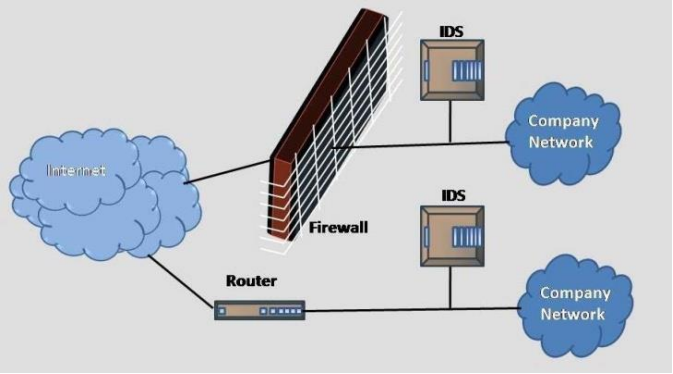

Figure 1: Typical intrusion detection system

It necessitates the use of dynamic technology that can track the system and detect criminal activity. As a consequence, a dynamic solution known as Intrusion Detection Technology is adopted to strengthen network stability. The intrusion monitoring device gathers online data from the network, tracks and analyses it, and divides it into natural and malicious operations before informing the system administrator [2].

Due to restricted scalability, adaptability, and authenticity, IDS is the field where Data mining is used extensively. Data is compiled from various sources in IDS, including network log data, host data, and so on. The analysis of data is too complicated, owing to the high level of network

traffic. This necessitates the use of IDS in conjunction with various data mining methods for intrusion detection.

This paper is organized as follows. Section 1 gives the Introduction. Section 2 discusses the literature survey: section 3 overviews the intrusion detection system and its classification. Section 4 gives various data mining techniques for IDS. Section 5 denotes the conclusion.

II. BACKGROUND STUDY

This section reviews related work on network-based data sets for intrusion detection.

A, A. H., & Sundarakantham, K. [1] In today's world, intrusion detection and prevention are important. Intrusion detection and intrusion protection are critical because our daily activities rely heavily on networks and information systems. Intrusion detection methods have utilized several techniques.

Abdulhammed, R. et al. [2] One of the most challenging problems in this area is attribute selection since the selected attributes impact the classification results.
Dang, Q.-V. [3] to improve lightweight intrusion detection methods by using an outlier detection algorithm as the active learning base learner, The algorithm does not necessitate a lot of processing power. As a consequence, it is ideal for low-power devices such as IoT devices or smartphones.

R, V., Alazab, M. et al. [5] To evaluate network and host-level operations, a hybrid intrusion detection warning system based on a highly scalable framework running on commodity hardware was proposed. For managing and analyzing very large scale data in real-time, the system used a distributed deep learning model with DNNs. The DNN model was chosen after a thorough evaluation of its results against traditional machine learning classifiers on a variety of IDS benchmark datasets.

Liu, J., & Chung, S. S. [6] The feature extraction phase aids feature selection by re-representing the original feature set in a more sensitive way to feature selection methods later on. To consider the combined effect of the features on the classification, feature selection methods can use the recursive feature elimination algorithm.

Waskle, S. [9] As the use of the internet by systems expands, security issues are becoming more prevalent. The proposed method effectively detects intruders over the internet. In contrast to previously used algorithms such as SVM, Nave Bayes, and Decision Tree, the proposed algorithm worked well. The proposed solution has the potential to increase identification rates and false error rates significantly.

III. NETWORK INTRUSION DETECTION SYSTEM

3.1 Common types of Intrusion Detection:

There is a wide spectrum of IDS, varying from antivirus software to hierarchical systems that monitor an entire backbone network's traffic. The most common classifications are network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS).

1. Network-Based (Network IDS):

Based on network traffic, network-based intrusion detection aims to detect unauthorized, illegal, and suspicious activity. A network IDS collects packets that traverse a network using a network tap, span port, or hub—the IDS system processes and flags any unusual traffic based on the collected data. The distinction between an intrusion detection system and an intrusion prevention system is that an intrusion detection system does not deliberately block network traffic.

2. Host-Based (HIDS):

Host-based intrusion detection, or HIDS, aims to detect unauthorized, illegal, or suspicious activity on a particular computer. In most cases, HIDS includes installing an agent on each device that controls and warns local OS and application operation. Unauthorized behaviour is detected by the configured agent using a combination of signatures, rules, and heuristics. A host IDS' function is largely passive, consisting of only gathering, identifying, recording, and alerting.

3. Physical (Physical IDS):

The act of detecting threats to physical structures is known as physical intrusion detection. Physical intrusion detection is often viewed as a collection of physical safeguards to protect the CIA. Physical intrusion detection systems are often used as prevention systems as well.

3.2 Classification of Intrusion Detection Based on Detection Approach:

It is also possible to classify IDS by detection approach:

1. Signature-based detection:

Misuse identification is another name for it. As a result, misuse detection is a signature-based IDS, in which intrusion detection is based on established attack behaviours, such as antivirus apps. Antivirus software compares the data to the virus's known code. In misuse detection, the history of identified malicious behaviour is stored in the dataset, and new instances are compared to the stored pattern of attacks to identify suspicious data.

2. Anomaly-based detection:

It's not the same as identifying misuse. The system administrator defines the baseline of normal data in-network data in a network, such as a network traffic load, protocol and packet size, etc. The Anomaly detector keeps track of new instances based on this baseline. The new instances are compared to the baseline, and any deviation from the baseline is reported as an intrusion. As a result, it's often referred to as a behavior-based intrusion detection system.

COMPARATIVE ANALYSIS OF SURVEY

Table 1: Evaluation of various authors views.

Paper Name	Methodology	Limitations
A, A. H., & Sundarakantham, K. [1]	SVM and Naïve Bayes Algorithm is Proposed	A distinct intrusion existence can steal or eliminate information from computer or network systems in a limited duration.
Dang, Q.-V. [3]	outlier detection algorithm	intrusion detection methods have limited
Chabathula, K. J., et al. [4]	Tree classification algorithms	the standard deviation limit
R, V., Alazab, M. et al. [5]	Deep neural network for intelligent intrusion system	Limitations and malicious attackers can gain unauthorized access to the system.
Waskle, S. et al. [9]	the detection of intruders over the internet efficiently	The detection rates and the false error rates are limited

IV. DISCUSSION

4.1 DATA MINING TECHNIQUES FOR INTRUSION DETECTION:

There are many data mining techniques for intrusion detection, such as frequent pattern mining, classification, clustering, mining data streams, etc. Let us see some of them here.

4.2 Classification:

Classification is the process of assigning a specific class to each dataset under consideration. For new cases, the words "usual" and "abnormal" apply to the use of a known structure. It can be used for both misuse detection and anomaly detection, but it's most widely used for the former. The datasets were classified into predetermined sets using classification. As compared to clustering, it is less effective at detecting intrusions. IDS employs several classification methods, including decision trees, naive Bayes classifiers, K-nearest neighbour classifiers, and support vector machines.

4.3 Decision Tree:

A decision tree is a recursive structure that looks like a tree and conveys classification rules. It splits the data according to attribute values using the divide and conquers process. The data is classified from root to leaf nodes, with each node representing an attribute and its value and each leaf node representing a data class mark. In the case of large datasets, tree-based classifiers perform best. Below are descriptions of various decision tree algorithms.

4.4 K-Nearest Neighbor:

It is one of the most basic classification methods. It computes the distance between different data points on the input vectors and assigns the unlabeled data point to its nearest neighbour's class. K is a critical parameter. If k=1, the object is allocated to the nearest neighbour's class. When K is high, it takes a long time to predict and impact accuracy by reducing the effect of noise.

4.5 Naive Bayes classifier:

Naive Bayes classifier is a probabilistic classifier. It predicts the class according to membership probability. To derive conditional probability, it analyzes the relationship between the independent and dependent variable.

4.6 Support Vector Machine

Support Vector Machines SVM is the most basic and popular machine learning method for classification and regression. This approach provides a series of training examples; each labelled as belonging to one of two categories. Using the Support Vector Machines algorithm, a model is constructed to predict whether a new example belongs to one of two categories.

V. CONCLUSION

In the computational intelligence community, machine learning for intrusion detection has gotten a lot of attention. Massive amounts of audit data must be processed in an intrusion detection algorithm to create new detection rules for a growing number of novel attacks in a high-speed network. To increase detection speed and accuracy, intrusion detection algorithms should consider the dynamic properties

of attack behaviours. Analyze a large volume of network datasets to improve detection accuracy; intrusion detection is becoming a popular machine learning research subject. Machine learning-based detection systems have the advantage of detecting or categorizing persistent features without any input from the environment. Learning-based detection systems have the disadvantage that if a sufficient amount of regular traffic data is not usable, the techniques' training becomes extremely difficult. Each method for implementing an intrusion detection system has its own set of benefits and drawbacks. It is clear from the discussion of the different methods' comparisons. As a result, deciding which approach to use to implement an intrusion detection system is difficult.

REFERENCES

[1] A, A. H., & Sundarakantham, K. (2019). Machine Learning-Based Intrusion Detection System. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). doi:10.1109/icoei.2019.8862784

[2] Abdulhammed, R., Faezipour, M., Abuzneid, A., & Alessa, A. (2018). Enhancing Wireless Intrusion Detection Using Machine Learning Classification with Reduced Attribute Sets. 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC). doi:10.1109/iwcmc.2018.8450479

[3] Dang, Q.-V. (2020). Active Learning for Intrusion Detection Systems. 2020 RIVF International Conference on Computing and Communication Technologies (RIVF). doi:10.1109/rivf48685.2020.9140751

[4] Chabathula, K. J., Jaidhar, C. D., & Ajay Kumara, M. A. (2015). Comparative study of Principal Component Analysis based Intrusion Detection approach using machine learning algorithms. 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). doi:10.1109/icscn.2015.7219853

[5] R, V., Alazab, M., KP, S., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. IEEE Access, 1–1. doi:10.1109/access.2019.2895334

[6] Liu, J., & Chung, S. S. (2019). Automatic Feature Extraction and Selection For Machine Learning Based Intrusion Detection. 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). doi:10.1109/smartworld-uic-atc-scalcom-iop-sci.2019.00254

[7] Subbulakshmi, T., BalaKrishnan, K., Shalinie, S. M., AnandKumar, D., GanapathiSubramanian, V., & Kannathal, K. (2011). Detection of DDoS attacks using Enhanced Support Vector Machines with real time generated dataset. 2011 Third International Conference on Advanced Computing. doi:10.1109/icoac.2011.6165212

[8] Manna, A., & Alkasassbeh, M. (2019). Detecting network anomalies using machine learning and SNMP-MIB dataset with IP group. 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS). doi:10.1109/ictcs.2019.8923043

[9] Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). doi:10.1109/icesc48915.2020.9155656

[10] Zwane, S., Tarwireyi, P., & Adigun, M. (2019). Ensemble Learning Approach for Flow-based Intrusion Detection System. 2019 IEEE AFRICON. doi:10.1109/africon46755.2019.9133979