IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review on Intrusion Detection System using Deep Learning

Ms. Tanushri Jain[1], Prof. Chetan Gupta[2]
jain.tanu.1995094@gmail.com[1], chetangupta.gupta1@gmail.com[2]
M.Tech Scholar, PG Dept. of CSE[1], Assistant Professor, PG Dept. of CSE[2]
SIRTS, Bhopal[1], SIRTS, Bhopal[2]

Abstract: **There is drastic increase in needs of networking and data sharing in today's world. Such globalization of increased information technology and development there exists need of network security. Firewalls may provide some level of security but they never alert administrator for upcoming attacks. In order to find such abnormal behavior of network packets there is need of reliable detection system for improvement of efficiency and accuracy. As in today's developing network environment there is threat of new type of attacks daily in the network. So, the network administration system is also needed to be updated regularly for up gradation of security level. One of the network packet monitoring system is Intrusion detection systems (IDS).There are many techniques in the literature for developing these defense systems. However, it is also important to examine the improvement of the datasets used to train and test these security systems. Enhanced datasets extend the detection capabilities of offline and online intrusion detection models. Standard datasets such as KDD 99 and NSL-KDD are obsolete and do not contain data on current attacks such as denial of service. Therefore, they are not suitable for evaluation. This article presents an in-depth analysis of IDS records and presents the challenges of IDS. This article also provides an overview of the deep learning approach that can be used to develop a better network intrusion detection system.**

Keywords: **Intrusion Detection, Security System, Deep learning, Attack.**

I. INTRODUCTION

Network security has become one of the most concerning problems for internet users and service providers with drastic increase in the internet usage [1]. A secure network is defined in terms of the protection of its software and hardware in contrast to different types of intrusions. A network is secured by applying a robust observation, analysis and defense mechanisms. As the world has become more connected over the Internet, computer networks are more prone to malicious attacks [2]. "Intrusion is an attempt to compromise CIA (Confidentiality, Integrity, Availability), or to bypass the security mechanisms of a computer or network" [3]. "Intrusion detection is the process of monitoring the events occurring in a computer system or network, and analyzing them for signs of intrusion" [4]. Intrusion detection is the process of an examining the events occur on a computer system or on a network are investigated and analyzed to detect signs of possible incidents that constitute violations or threaten to violate computer security policies, use policies acceptable security practices or standard [5].

An intrusion detection system (IDS) is software that automates the intrusion detection process. Intrusion prevention is the process of detecting intrusions and attempting to block potential identified incidents. NIDS is one of the main tools used to report network attacks. An IDS system which uses network behavior is called as NIDS. The network behaviors are collected using network equipment via mirroring by networking devices, such as switches, routers, and network taps and analyzed in order to identify attacks and possible threats concealed within in network traffic.

An IDS system which uses system activities in the form of various log files running on the local host computer in order to detect attacks is called as HIDS. The log files are collected via local sensors. While NIDS inspects each packet contents in network traffic flows, HIDS relies on the information of log files which includes sensors logs, system logs, software logs, file systems, disk resources, users account information and others of each system. Many organizations use a hybrid of both NIDS and HIDS.

A network intrusion detection system (NIDS) monitors the network traffic by identifying suspicious activity, which may represent an attack or illegal access. NIDS are tools which implement such mechanisms so as to protect a network from intrusions which may be from within the network or from outside the network [6][7]. These systems observe the incoming and also outgoing traffic of a network, perform analysis periodically and report when an intrusion is detected, as shown in Fig. 1.

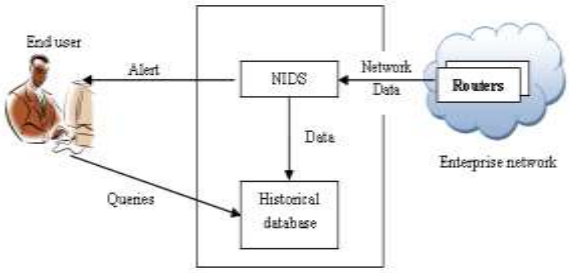

Fig. 1. Network Intrusion Detection System

II. INTRUSION DETECTION DATASETS

Evaluation records play an important role in validating an IDS approach by allowing us to evaluate the ability of the proposed method to detect intrusive behavior. The datasets used for analyzing network packets in commercial products are not readily available for data protection reasons. However, there are publicly available records such as DARPA, KDD, NSL-KDD and ADFA-LD which are commonly used as a reference. The existing datasets used to create and compare IDSs are explained in this section with their functions and restrictions.

A. KDD Cup 99 Dataset

The first attempts to create an IDS record were made in 1998 by the Defense Advanced Research Project Agency (DARPA) and created the Knowledge Discovery and Data Mining (KDD) record. In 1998, DARPA introduced a programmer to the MIT Lincoln labs to provide a complete and realistic IDS benchmarking environment (MIT Lincoln Laboratory, 1999). Although this dataset was an important contribution to IDS research, its accuracy and ability to take actual conditions into account has been widely criticized.

These datasets were collected using multiple computers connected to the Internet to model a small American air base with limited personnel. Network packets and host log files were collected. Lincoln Labs created an experimental test environment to get a snapshot of the 2-month TCP packet for a Local Area Network (LAN) using a typical US Air Force LAN. They modeled the LAN as if it were a real Air Force environment, but nested it with multiple simulated interventions. The network packets collected were approximately four gigabytes in size and contained approximately 4,900,000 datasets. The 2-week test data contained approximately 2 million connection records, each with 41 characteristics and were classified as normal or abnormal. The extracted data are a series of TCP sessions that begin and end at specific times, among which the data flows from and to a source IP address to a destination IP address that contains a variety of attacks that occur in an environment of military network. The 1998 DARPA dataset was used as a basis for deriving the KDD Cup99 dataset, which was used in the third international competition for knowledge discovery and data mining tools (KDD, 1999).

The 1999 KDD Cup dataset was used for the third international competition for the discovery of knowledge and data mining tools. Each connection instance is described by 41 attributes (38 continuous or discrete numeric attributes and 3 symbolic attributes). Each instance is called a normal attack or a specific type of attack. These attacks fall into one of four categories: DoS, Probe, U2R and R2L. The 1999 KDD Cup provided training and test data sets, called 10% KDD or correct data sets. The 10% KDD dataset contains 22 types of attacks, while the correct dataset contains the same 22 types of attacks and 17 types of additional attacks. These records are obsolete because they do not contain any records of recent malware attacks. For example, the behavior of attackers differs between different network topologies, operating systems and different criminal software and toolkits. Nonetheless, KDD99 continues to be used as a reference within the IDS research community and is currently still used by researchers.

B. NSL-KDD Dataset

NSL-KDD is a public dataset, which has been developed from the earlier KDD-99 dataset. Statistical analysis of the cup99 dataset resulted in significant problems that significantly affect the accuracy of intrusion detection and lead to a misleading assessment of AIDS. The main problem with registering KDD is the large number of duplicate packages. Tavallaee et al. [8] analyzed the KDD training and test sets and found that approximately 78% and 75% of network packets are duplicated in training and test data sets. This huge number of duplicate instances in the training set would affect the machine learning methods aimed at normal instances, preventing them from learning from irregular instances that normally damage the computer system. Tavallaee et al. [8] created the 2009 NSL-KDD dataset from the KDD Cup'99 dataset to solve the above problems by eliminating duplicate datasets. The NSL-KDD train data record contains 125,973 data records and the test data record contains 22,544 data records. The size of the NSL-KDD record is sufficient to facilitate the use of the entire NSL-KDD record without the need for a random sample. This has led to consistent and comparable results from various research projects. The NSL_KDD dataset includes 22 training intrusion attacks and 41 attributes (i.e. characteristics). In this dataset, 21 attributes refer to the connection itself and 19 attributes describe the type of connection within the same host.

C. CICIDS 2017 Dataset

The CICIDS2017 dataset contains both anomaly behavior and details about new malware attacks: Brute Force FTP, Brute Force SSH, DoS, Heart bleed, Web Attack, Infiltration, Botnet and DDoS. This record is identified by date/time, source and destination IP addresses, origin and destination ports, protocols and attacks. A complete network topology has been configured for the acquisition of this dataset, which contains modems, firewalls, switches, routers and nodes with various operating systems (Microsoft Windows (such as Windows 10, Windows 8, Windows 7 and WindowsXP), Apple MacOS iOS and contains an open source Linux operating system). This dataset contains 80 network flow functions from recorded network traffic.

D. UNSW-NB 15 Dataset

The raw network packets of the UNSW-NB 15 dataset were created by the IXIA Perfect Storm tool in the Australian Center for Cyber security (ACCS), Cyber Range Lab to generate a mixture of normal and anomaly behavior. The Tcpdump tool is used to acquire 100 GB of raw data traffic (e.g. Pcap files). This dataset contains nine types

of attacks: Fuzzer, Analysis, Backdoor, DoS, Exploits, Generic, Reconnaissance, Shell code and Worms. Argus, Bro-IDS tools and twelve developed algorithms are used to generate a total of 49 class-labeled features [9].

III. CHALLENGES DURING DEVELOPMENT OF IDS

Some challenges of IDS are discussed as below:

A. Challenges related to nature of datasets

Two main categories of challenges arise during the development of NIDSs to obscure future attacks. The first challenge is suitable feature selection. The second challenge is inaccessibility of labeled traffic datasets. Following challenges are related to nature of dataset:

- The imbalanced and diverse nature of the datasets.
- Unavailability of labeled traffic dataset from real networks.
- Misclassification of targeted input pattern.

B. Challenges related to data processing increase

The use of new technologies in field of network communication leads to lower imperfection rates and therefore generates a huge amount of network data. The following challenges concern data processing.

- The processing of data is increased.
- Difficulty to process big data.
- Long time processing or computational complexity is high.
- Difficulty to process large network data for packet classification as there exists millions of packets.

C. Challenges related to security

Challenges in security are divided into two categories. Firstly, the security of every machine presented currently to the Internet can be compromised and attacked. External attacks constantly threaten important data. Hackers discover new methods to steal or damage valuable data in every organization every day. Secondly, security demand is crucial to many companies and organizations that depend on a database to safeguard sensitive data. The value of certain data is worth millions. Thus, strong data protection must be guaranteed.

D. Challenges related to growth of new attacks

Smartphone malware is currently used in daily life activities, such as entertainment, controlling smart homes and paying bills. Smart phones have demonstrated a considerable increase in growth rate given their mobility and ever-expanding capabilities. Android is an ideal platform for legitimate developers and attackers creating malware given its large market share and room for development.

IV. LITERATURE REVIEW

Liang et al. [1] used a hybrid positioning strategy based on a multi-agent system for the intrusion detection system. This system includes a data acquisition module, a data management module, an analysis module and a response module. In this study, an algorithm for a deep neural network for intruder detection is used to implement the analysis module. The results show the effectiveness of deep learning algorithms for detecting transport-layer attacks.

Moustafa et al. [10] generated a UNSW-NB15 dataset for intruder detection. This dataset contains nine types of modern attack modes and new normal traffic patterns. It contains 49 attributes that include host-flow and network packet control to distinguish between normal and abnormal observations. In this article, we demonstrate the complexity of the UNSW-NB15 dataset in three ways. First, the statistical analysis of observations and attributes is explained. Second, the study of characteristic correlations is provided. Third, five existing classifiers are used to assess the complexity in terms of accuracy and false alarms rate (FAR), then the results are compared with the KDD99 dataset. Experimental results show that UNSW-NB15 is more complex than KDD99 and is considered a new reference dataset for the evaluation of NIDS.

Zhao et al. [11] proposed an intrusion detection method using a deep belief network (DBN) and a probabilistic neural network (PNN). First, the raw data is converted into small data, while the essential attributes of the raw data are preserved using DBN's nonlinear learning ability. Second, a swarm of particle optimization algorithms are used to improve learning performance to optimize the number of nodes with hidden levels per level. Subsequently, PNN is used to classify low dimensional data.

ChuanlongYin et al. [12] examined the model's performance in binary classification and the classification of different classes, as well as the number of neurons and the different effects of the learning rate on the performance of the proposed model. We compare it with those of J48, the artificial neural network, the random forest, the support vector machine and other machine learning methods suggested by previous researchers for the reference dataset. Experimental results show that RNN-IDS is very suitable for modeling a classification model with great precision and that its performance is superior to conventional classification methods for machine learning in binary and multi-class classification.

Yuan et al. [14] proposed a DDoS attack detection approach based on deep learning (Deep Defense). The deep learning approach can automatically extract high-level functions from low-level functions and achieve powerful representation and conclusion. A recurring deep neural network project is proposed to learn patterns from network traffic sequences and to track network attack activity. Experimental results show that the model works better than traditional machine learning models, as it reduces the error rate from 7.517% to 2.103% compared to conventional machine learning methods in the larger dataset.

Ma T et al. [19] proposed a new approach called KDSVM, which took advantage of k-mean techniques and the advantage of learning functionality with a deep neural network (DNN) model and a support vector machine (SVM) classifier to detect intrusion networks. KDSVM consists of two phases. In the first step, the data set is divided into k subsets as a function of each sampling distance from the cluster centers of the k-means approach and, in the second step, the test data set is far from the same cluster center and entered in the DNN model with SVM.

Feng et al. [20] has introduced a machine learning based data classification algorithm that is applied to intrusion detection in the network. The basic activity is to classify network activities in the network protocol as connection records such as normal or abnormal, minimizing classification errors. Although various classification models have been developed for network intrusion detection, each has its own strengths and weaknesses, including vector machine methods.

Ali et al. [21] proposed a hybrid machine learning technique for detecting network intrusions, which is based on a combination of K-medium clusters and Sequential Minimal Optimization (SMO) classification. A hybrid approach is introduced to reduce the rate of false alarms and false alarms, improve the detection rate and detect attackers on day zero. The NSL-KDD dataset was used in the proposed technique. Classification was performed using Sequential Minimal Optimization.

Laftah et al. [22] proposed a modified K-mean algorithm to create a high-quality training dataset that greatly improves the performance of classifiers. The modified K-mean is used to create new small training datasets that represent the complete set of original training data, significantly reduce the time spent training classifiers and improve the performance of the intrusion detection system.

Table I: Contribution of Machine learning in field of IDS

Author Name	**Approach Used**	**Average Detection Rate or Accuracy**
Liang et al. [1]	For Internet of Things based on a Machine Learning approach.	97%
Feng et al. [20]	SVM and Clustering based on Self-Organized Ant Colony Network.	94.86 %
Saad Mohamed et al. [21]	K-means clustering and Sequential Minimal Optimization (SMO) classification.	97.36%
Wathiq Laftah Al-Yaseen et al. [22]	SVM and Extreme machine learning technique.	95.75%
Manjula C. Belavagi et al. [23]	Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest.	About 99 %

V. DEEP LEARNING BASED IDS

Deep learning is an emerging trend in the area of machine learning. It is sub-field of machine learning in artificial neural networks. Using deep learning approach in the application area, we can process on large amount of items in order to be trained. Process is placed on millions of data points. Deep learning is learns features from the data. If large amount of data is available, it can reduce the performance of system. For achieving better accuracy in terms of performance deep learning is well suited learning mechanism. Some research works related to deep learning in field of intrusion detection are summarized below in table II.

Table II: Contribution of Deep Learning in field of IDS

Technique	**Attack Types**	**Metrices**
Recurrent Neural Network [11][12][13]	DoS, R2L, U2R and probe	Detection Rate and False Alarm Rate
	DoS, R2L, U2R and probe	Detection Rate and False Alarm Rate
	HTTP Web, unknown TCP, secure web, misc application, SMTP, IMAP, Flowgen, ICMP, DNS, IRC	Error rate, Accuracy, Precision, Recall, F1-score and AUC
Deep Belief Network[14]	Android malware	Precision, Recall and F1-score
Stacked auto encoder[15]	DoS, R2L, U2R and probe	Detection rate and false alarm rate
Stacked denoising auto encoders [16]	PC malware	Accuracy, Precision, Recall and F1-score
Convolutional Neural-Learning Classifier System (CN-LCS) [17]	Abnormal queries	Accuracy
Deep Neural Network with Support Vector Machine and Clustering Technique [18]	Dos, Probe, U2R, R2L	Error rate, Accuracy, Recall

VI. PROBLEM DOMAIN

It is well known that anomaly-based IDS suffer from the high rate of false alarms. Continuous efforts are being made to reduce the high false positive rate. We believe that intrusion detection is a data analysis process and can be studied as a problem of classifying data correctly. From this standpoint, it can also be observed that any classification scheme is as good as the data presented to it as input. Cleaner the data, higher accurate results are likely to be obtained. From anomaly-based IDS point of view, it implies that if we can extract features that demarcate normal data from abnormal one properly, false positive rate can be reduced to a great extent. On the similar lines, we observe that most of the data mining and machine learning based methods in intrusion detection make use of well-known tools and techniques. It may turn out that these general techniques are not very effective in classifying data as normal or abnormal with very high accuracy. There is a need to customize those techniques according to the requirement of intrusion detection. Apart from the problems mentioned above, the fast detection of attacks remains one of the focal points to be worried about. With the present complexity and variety of attacks, we need a huge amount of data to analyze and produce results. But larger the amount of data, longer the time to analyze it, which delays the detection of attacks. An IDS will be of more use if it can trigger an alarm early enough to reduce the damage that an ongoing attack can do. Thus, there is a need to make IDS as fast as to operate on-line. It is believed that this can be achieved if we can reduce the data, to be analyzed, without degrading its quality.

VII. CONCLUSION

Information security has become a legitimate concern for organizations and computer users due to the growing trust in computers and electronic transactions. Various techniques are used to ensure a company's security against threats or attacks. On the other hand, attackers are discovering new techniques and ways to violate these security guidelines. The main types of IDS technologies - network-based, wireless and host-based offer substantially different functions. This paper reviews and analyses the research area for intrusion detection systems (IDSs) based on deep learning (DL) techniques into a coherent taxonomy and identifies the gap in this pivotal research area.

REFERENCES

[1] Chao Liang, Bharanidharan Shanmugam, Sami Azam, Mirjam Jonkman, Friso De Boer, Ganthan Narayansamy, "Intrusion Detection System for Internet of Thingsbased on a Machine Learning approach", International Conference on Vision Towards Emerging Trends in Communication and Networking, IEEE, 2019.

[2] Amudhavel, J., Brindha, V., Anantharaj, B., Karthikeyan, "A survey on intrusion detection system: State of the art review". *Indian J. Sci. Technol.* Vol. 9, pp. 1–9, 2016

[3] Patel, A., Taghavi, M., Bakhtiyari, K., Júnior, J.C. "An intrusion detection and prevention system in cloud computing: a systematic review". *J. Netw. Comput. Appl.* Vol. 36, pp. 25–41, 2013.

[4] S. Ashoor and S. Gore, "Importance of Intrusion Detection system (IDS)", *International Journal of Scientific and Engineering Research*, vol. 2, issue 1, pp. 1-4, 2011.

[5] Y. Farhaoui and A. Asimi, "Performance assessment of tools of the intrusion detection/prevention systems," *International Journal of Computer Science and Information Security*, vol. 10, issue 1, pp. 1-7, 2012.

[6] L. N. De Castro, "An introduction to the artificial immune systems," *in Handbook of Natural Computing*, pp. 1575–1597, 2012.

[7] Y. Farhaoui and A. Asimi, "Performance method of assessment of the intrusion detection and prevention systems," *International Journal of Engineering Science and Technology*, vol. 3, issue. 7, 2011.

[8] K. Boukhdir, A. Boualam, S. Tallaland H. Medromi, and S. Benhadou, "Conception, design and implementation of secured uav combining multi-agent systems and ubiquitous lightweight idps (intrusion detection and prevention system)," *International Journal on Engineering Applications*, vol. 3, issue 1, pp. 1– 5, 2015.

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE symposium on computational intelligence for security and defense applications, 2009, pp. 1–6.

[10] Nour Moustafa & Jill Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set", Information Security Journal: A Global Perspective, 2015.

[11] Zhao, G.; Zhang, C.; Zheng, L. "Intrusion detection using deep belief network and probabilistic neuralnetwork", In Proceedings of the 2017 IEEE International Conference on Computational Science andEngineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),Guangzhou, China, 21–24 July 2017; Volume 1, pp. 639–642.

[12] Yin C et al , "A deep learning approach for intrusion detection using recurrent neural networks", IEEE Access 5:21954–2196, 2017.

[13] Kim J, Kim H, "Applying recurrent neural network to intrusion detection with hessian free optimization", International workshop on information security applications. Springer, 2015.

[14] Yuan X, Li C, Li X, "DeepDefense: identifying DDoS attack via deep learning", IEEE international conference on smart computing (SMARTCOMP), 2017.

[15] Wang Z et al, "Droiddeeplearner: identifying android malware using deep learning", Sarnoff symposium. IEEE, 2016.

[16] Jing L, Bin W, "Network intrusion detection method based on relevance deep learning", international conference on intelligent transportation, big data & smart city (ICITBS). IEEE, 2016.

[17] Bu S-J, Cho S-B, "A hybrid system of deep learning and learning classifier system for database intrusion detection", International conference on hybrid artificial intelligence systems, SpringeR, 2017.

[18] Kim J et al, "Long short-term memory recurrent neural network classifier for intrusion detection", International conference on platform technology and service (PlatCon), IEEE, 2016.

[19] Ma T et al, "A hybrid methodologies for intrusion detection based deep neural network with support vector machine and clustering technique", International conference on frontier computing. Springer, 2016.

[20] Feng, W., "Mining network data for intrusion detection through combining SVMs with ant colony networks", Future Gener. Comput. Syst., 2014, 37, 127–140.

[21] Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique, International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

[22] Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman ,Mohd Zakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", International Journal in Expert Systems With Applications, Elsevier, 2017.

[23] Manjula C. Belavagi and Balachandra Muniyal, Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science, Elsevier, 2016.