International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 07, Issue: 04 (July-August, 2021)

Review of Cyber Intrusion Detection System Based on Different Approach of Feature Selection and Classification

Moti Kumari, Prof. Monika Raghuvanshi
M. Tech Research scholar Department of CSE
Bhabha Engineering Research Institute, MP, Bhopal, India
motikumari583@gmail.com, monipriya@gmail.com

Abstract: The malicious activity is increasing day to day in the cyber world. The increased activity defamed the reputation of data and organisation. For the controlling of malicious activity, various hardware and software level system is implemented. Nevertheless, the system's limitation is detecting dynamic nature-based intruders and suspicious events over a cyber network. This paper presents a review of intrusion detection based on different features of intrusion. The Intrusion detection methods are two different modes, such as host-based intrusion detection and network-based intrusion detection. The host-based intrusion detection system applied signature-based methods, and the algorithm's efficiency is high compared to the anomaly-based intrusion detection system. For the validation of the intrusion detection system applied different datasets such as KDDCUP99, DARPA98, NSL-KDD. This survey paper presents a taxonomy of new IDS, a comprehensive review of notable recent works, and an overview of the datasets commonly used for evaluation purposes. It also presents evasion techniques attackers use to avoid detection and discusses future research challenges to counter such techniques to make computer systems more secure.

Keywords: IDS, Malicious activity, anomaly detection, KDDCUP99, Machine Learning, features, Detection.

I. Introduction

The growth of malicious software over the internet poses a challenging task for the intrusion detection system. The attacker developed various complex codes and changed the normal traffic feature sequence [1]. In addition, there has been an increase in security threats such as zero-day attacks designed to target internet users. Therefore, computer security has become essential as information technology has become part of our daily lives. As a result, various countries such as Australia and the US have been significantly impacted by the zero-day attacks [2,3]. However, the new generation of malware has become more ambitious and targets the banks themselves, sometimes trying to take millions of dollars in one attack. The IDS provides one of the most promising paths towards network robustness. IDS are used to detect several types of malicious activities that can violate the rules and trust of the host systems. IDS can further classify as Anomaly-based intrusion detection and Signature-based intrusion detection. Anomaly-based IDS determines normal network behaviour like bandwidth range, types of protocols, ports, and a device used to connect each other, and alerts sent to the network administrator or user when inconsistent traffic is detected [4,5,6,7,8].

On the other hand, Signature-based IDS monitors packets in the network and compares them with pre-configured and preidentified attack behaviours, called signatures—the dataset created by combining the information in the system logs. The major challenges in the existing KDD'99 based IDS are that the data set has many attributes, increasing the system's processing time. The accuracy of the existing classifiers like C4.5, Random forest is also low. The KDD'99 dataset consists of 42 features [9,8,10,11,12]. So, it requires more time to process, and it takes longer to detect the known attacks.

The fast detection process of intrusion direction focuses on the selection of features and classification. The process of features reduction enhanced the performance of the intrusion detection system. Now day used various features reduction algorithms are used for static as well as dynamic features reduction. The feature reduction technique behaves in dual mode. The reduction of features cannot have fixed how many features are reduced to detect intrusion better. The reduction process used the PCA method. This method is a static reduction technique, reduces the only fixed number of attributes. The fixed number of feature reduction processes does not justify the value of the feature. It directly reduces the feature.

Regarding computational time, feature reduction is also an important aspect, and the reduced feature increases the processing of detection ratio [16,17,18]. Many methods have been proposed in the last decades on the designs of IDSs based on feature reduction techniques. The extraction of features is a major issue in intrusion detection. The intruder file features are mixed categories.

Some features are numeric, some are alphabetical, and some are constant to extract features using the transformation technique. The transformation technique converts all features in the same nature and property [19,20]. The data mining gives the data transformation algorithm is called a min-max algorithm. The min-max algorithm converts all features in numeric data in a range of (0-1). 0 is the value of the minimum feature, and 1 is the maximum feature value. The role of the classification algorithm is very important in the detection of dynamic nature-based malicious software agents. Machine learning

provides various classification algorithms for the detection of intrusion. The rest of the paper is organised as in section II related work in section III—problem formulation. In section IV approach is applied and finally discuss the conclusion & future work.

II. Related Work

The challenge of cyber-attacks increases in our daily life in different formats. The developments of algorithms and models by the continuous approach of different authors and research scholars. Here describes the contribution of different authors in the area of intrusion detection.

In [1] Author do not manually design the features of the flow but directly extract the raw data information of the flow for analysis. In addition, author first discussed a new network intrusion detection model named the deep hierarchical network, which integrates the improved LeNet-5 and LSTM neural network structures while learning the spatial and temporal features of flow. By designing a reasonable network cascading method, the author can simultaneously train their discussed hierarchical network instead of training two networks separately. In this paper, the author uses the CICIDS2017 dataset and the CTU dataset. The number and types of flow in these two datasets are large, and the attack types are relatively new. The experimental results show that the performance of the discussed hierarchical network model is significantly better than other network intrusion detection models, which can achieve the best detection accuracy. In [2], Information and Communication Technology (ICT) greatly impacts social well-being, economic growth, and national security in today's world. Generally, ICT includes computers, mobile communication devices and networks. A group of people also embraces ICT with malicious intent, known as network intruders, cybercriminals. Confronting these detrimental cyber activities is one of the international priorities and an important research area. Anomaly detection is an important data analysis task that is useful for identifying network intrusions. In [3] author constructs an IDS model with a deep learning approach. The author applies Long Short-Term Memory (LSTM) architecture to a Recurrent Neural Network (RNN) and trains the IDS model using KDD Cup 1999 dataset. Through the performance test, the author confirms that the deep learning approach is effective for IDS. In [4] Author discussed and empirically evaluate a novel network-based anomaly detection method that extracts behaviour snapshots of the network and uses deep autoencoders to detect anomalous network traffic emanating from compromised IoT devices. The author infected nine commercial IoT devices in their lab with two of the most widely known IoT-based botnets, Mirai and BASHLITE. their evaluation results demonstrated their discussed method's ability to accurately and instantly detect the attacks as they were being launched from the compromised IoT devices which were part of a botnet. In [5], this paper aims to discover the principles of designing effective ConvNet architectures for video action recognition and learn these models given limited training samples. their first contribution is temporal segment network (TSN), a novel framework for video-based action recognition. which is based on the idea of long-range temporal structure modelling. It combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning using the whole action video. The other contribution is their study on a series of good practices in learning ConvNets on video data with the help of a temporal segment network. In [6] author present a survey of IDS research efforts for IoT. their objective is to identify leading trends, open issues, and future research possibilities. The author classified the IDSs discussed in the literature according to the following attributes: detection method, IDS placement strategy, security threat and validation strategy. The author also discussed the different possibilities for each attribute, detailing aspects of works that either discussed specific IDS schemes for IoT or developed attack detection strategies for IoT threats that might be embedded in IDSs. In [7], Cybersecurity continues to be a serious issue for any sector in cyberspace as the number of security breaches is increasing from time to time. It is known that thousands of zero-day attacks are continuously emerging because of various protocols, mainly from the Internet of Things (IoT). Most of these attacks are small variants of previously known cyber-attacks. It indicates that even advanced mechanisms such as traditional machine learning systems face the difficulty of detecting these small mutants of attacks over time. On the other hand, the success of deep learning (DL) in various big data fields has drawn several interests in cybersecurity fields. The application of DL has been practical because of the improvement in CPU and neural network algorithms aspects. In [8] Author discussed an application-aware traffic control scheme, in which both network states and traffic behaviours are exploited cooperatively. Deep Packet Inspection (DPI) is introduced into the SDN controller. Meanwhile, a packet classification and behaviour matching mechanism is designed to exchange information between components, and a publish/subscribe based middleware is designed. Besides, mathematical models for analysing network throughput and latency are established. Simulation results show that the discussed scheme can improve throughput and reduce the latency time of end-to-end communication. In [9] author discusses some widely used deep learning architectures and their practical applications. An up-to-date overview is provided on four deep learning architectures: autoencoder, convolutional neural network, deep belief network, and restricted Boltzmann machine. Different types of deep neural networks are surveyed, and recent progress is summarised. Applications of deep learning techniques in some selected areas (speech

recognition, pattern recognition and computer vision). In [10] author present a multi-step outlier-based approach for the detection of anomalies in network-wide traffic. The author identifies a subset of relevant traffic features and uses it during clustering and anomaly detection. To support outlier-based network anomaly identification, the author uses the following modules: mutual information and generalised entropy-based feature selection technique to select a relevant non-redundant subset of features, a tree-based clustering technique to generate a set of reference points and an outlier score function to rank incoming network traffic to identify anomalies. The author also designs a fast-distributed feature extraction and data preparation framework to extract features from raw network-wide traffic. The author evaluates their approach in terms of detection rate, false-positive rate, precision, recall and F -measure using several high dimensional synthetic and real-world datasets and finds the performance superior to competing algorithms. In [11] author discussed zeroth-order optimisation (ZOO) based attacks to directly estimate the gradients of the targeted DNN for generating adversarial examples. The author uses zeroth order stochastic coordinate descent, dimension reduction, hierarchical attack, and importance sampling techniques to attack black-box models efficiently. By exploiting zeroth-order optimisation, improved attacks to the targeted DNN can be accomplished, sparing the need for training substitute models and avoiding the loss in attack transferability. Experimental results on MNIST, CIFAR10 and ImageNet show that the discussed ZOO attack is as effective as the state-of-the-art white-box attack and significantly outperforms existing black-box attacks via substitute models. In [12] author introduces a defensive mechanism called defensive distillation to reduce the effectiveness of adversarial samples on DNNs. The author analytically investigates the generalizability and robustness properties granted using defensive distillation when training DNNs. The author also empirically study the effectiveness of their defence mechanisms on two DNNs placed in adversarial settings. The study shows that defensive distillation can reduce the effectiveness of sample creation from 95% to less than 0.5% on a studied DNN. Such dramatic gains can be explained by the fact that distillation leads gradients used in adversarial sample creation to be reduced by a factor of 1030. The author also finds that distillation increases the average minimum number of features that need to be modified to create adversarial samples by about 800% on one of the DNNs authors tested. In [13] Author discussed a distributed anomaly detection system using hierarchical temporal memory (HTM) to enhance the security of a vehicular controller area network bus. The HTM model can predict the flow data in real-time, which depends on the state of the previous learning. In addition, the author improved the abnormal score mechanism to evaluate the prediction. Author manually synthesised field modification and replay attack in the data field. Compared with recurrent neural networks and hidden Markov model detection models, the results show that the distributed anomaly detection system based on HTM networks achieves better performance in the area under the receiver operating characteristic curve score, precision, and recall. In [14] Author discussed a novel network intrusion model by stacking dilated convolutional autoencoders and evaluate their method on two new intrusion detection datasets. Several experiments were carried out to check the effectiveness of their approach. The comparative experimental results demonstrate that the discussed model can achieve considerably high performance, which meets the demand of high accuracy and adaptability of network intrusion detection systems (NIDS). It is quite potent and promising to apply their model in large-scale and real-world network environments. In [15], the author discussed associating the features from the static analysis with features from dynamic analysis of Android apps and characterising malware using deep learning techniques. Author implement an online deep-learning-based Android malware detection engine (Droid Detector) that can automatically detect whether an app is malware or not. With thousands of Android apps, the author thoroughly test Droid Detector and perform an in-depth analysis on the features that deep learning essentially exploits to characterise malware. The results show that deep learning is suitable for characterising Android malware and is especially effective with more training data. Droid Detector can achieve 96.76% detection accuracy, which outperforms traditional machine learning techniques. An evaluation of ten popular anti-virus software's demonstrates the urgency of advancing their capabilities in Android malware detection.

III. Problem Formulation

From the last decade of the internet, technology faced a problem of security threats and attacks. Security threats and attacks are categorised into two different domains. One is conventional domain such as signature and pattern-based attack. On the other side are dynamic attacks and threats. Conventional attacks are easily predicted and detected. But the dynamic attacks detection is very difficult due to a large number of attributes of network files. In the scenario of intrusion detection, various feature reduction methods are used for the fast detection process. In the journey of feature reduction used PCA algorithm, optimisation algorithm and some neural network models are used[6,7,8]. The reduction of features in network-based operation is a very critical issue. Some authors suggested the reduction cum classification technique for the intrusion detection system. In the process of review study, various research and journal papers related to feature reduction and detection of intrusion. The reduction of features is a very challenging task. The major issue is how many features is reducing for the process of detection. Some

authors suggested that the maximum number of reduces features is 10 to 15. Here discuss some problems related to feature reduction [19,20].

1. The processing of mixed categories of feature attributes is very difficult
2. Some feature attribute involves in both types of attacks.
3. The new attribute of attack passes as a normal attribute
4. The PCA and optimisation algorithm only reduces repeated features
5. The prediction of dynamic features as dual-mode.

The processing of feature and description of feature discuss in table 1,2 and 3 according to their description and data type[7,8]

Table 1 list of basic features of TCP connection

Feature name	Description	Type
hot	number of "hot" indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of ``compromised'' conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
num_root	number of ``root'' accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an FTP session	continuous
is_hot_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete

Table 2: Content features within a connection suggested by domain knowledge.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN'' errors	continuous
rerror_rate	% of connections that have ``REJ'' errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
srv_serror_rate	% of connections that have ``SYN'' errors	continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Table 3: Traffic features computed using a two-second time window.

Feature name	Description	Type
count	number of connections to the same host as the current connection in the past two seconds	continuous
serror_rate	% of connections that have ``SYN'' errors	continuous
rerror_rate	% of connections that have ``REJ'' errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
srv_serror_rate	% of connections that have ``SYN'' errors	continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

IV. Approach for Anomaly Detection

The efficiency of detection algorithms depends on the selection of features and optimisation of features. The optimised features reduce the traffic time of label categorisation of normal and abnormal traffic. The process of optimisation minimises the unwanted features for the process of classification. There are various optimisation algorithms such as genetic algorithm, particle swarm optimisation, and many more bio-inspired feature selection algorithms. Almost in the case of anomaly intrusion detection, the nature of features attribute is discrete and continuous. Some features attribute is ambiguous nature, and the optimisation process of features is very difficult. The extension works proposed hybrid feature selector methods for feature selection methods. The hybrid feature selector methods are based on plant grow optimisation (PGO) and glow-worm swarm optimisation (GSO).

V. Conclusion

This paper presents a review of the intrusion detection system. The reported survey of intrusion detection focuses on the area of work as feature optimisation and feature selection. The feature selection and optimisation process reduce cyber traffic overhead, and the detection process is improved. Various authors applied the mutual information-based feature selection methods, and the MI methods estimate the entropy of features. The entropy-based feature selection applied the statical based classifier for intrusion detection. Despite feature selection methods, various authors suggested linear sequence approaches to cover the payload based on the signature-based pattern. The classification algorithms are also essential assets of the intrusion detection system. Machine learning provides various classification algorithms such as supervised, unsupervised and reinforced learning. Some authors also applied the hybrid classification algorithm. The process of detection concludes on feature-based intrusion detection methods is better than another approach of intrusion detection.

REFERENCES

[1] Marteau, Pierre-Francois. "Sequence covering for efficient host-based intrusion detection." IEEE Transactions on Information Forensics and Security 14, no. 4 (2018): 994-1006.

[2] Tengl, Shaohua, Zhenhua Zhang, Luyao Teng, Wei Zhang, Haibin Zhu, Xiaozhao Fang, and Lunke Fei. "A collaborative intrusion detection model using a novel optimal weight strategy based on genetic algorithm for ensemble classifier." In 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), pp. 761-766. IEEE, 2018.

[3] Zhang, Qiangyi, Yanpeng Qu, and Ansheng Deng. "Network Intrusion Detection Using Kernel-based Fuzzy-rough Feature Selection." In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-6. IEEE, 2018.

[4] Abdul hammed, Razan, Miad Faezipour, Abdelshakour Abuzneid, and Arafat AbuMallouh. "Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic." IEEE sensors letter 3, no. 1 (2018): 1-4.

[5] Zhang, Hongpo, Chase Q. Wu, Shan Gao, Zongmin Wang, Yuxiao Xu, and Yongpeng Liu. "An effective deep learning-based scheme for network intrusion detection." In 2018 24th International Conference on Pattern Recognition (ICPR), pp. 682-687. IEEE, 2018.

[6] Hongwei, Ding, and Wan Liang. "Research on Intrusion Detection Based on KPCA-BP Neural Network." In 2018 IEEE 18th International Conference on Communication Technology (ICCT), pp. 911-915. IEEE, 2018.

[7] Marteau, Pierre-François. "Sequence covering similarity for symbolic sequence comparison." arXiv preprint arXiv:1801.07013 (2018).

[8] Deshpande, Prachi, Subhash Chander Sharma, Sateesh K. Peddoju, and S. Junaid. "HIDS: A host-based intrusion detection system for the cloud computing environment." International Journal of System Assurance Engineering and Management 9, no. 3 (2018): 567-576.

[9] da Costa, Kelton AP, João P. Papa, Celso O. Lisboa, Roberto Munoz, and Victor Hugo C. de Albuquerque. "Internet of Things: A survey on machine learning-based intrusion detection approaches." Computer Networks 151 (2019): 147-157.

[10] Zegeye, Wondimu K., Richard A. Dean, and Farzad Moazzami. "Multi-layer has hidden Markov model-based intrusion detection system." Machine Learning and Knowledge Extraction 1, no. 1 (2019): 265-286.

[11] Azad, Chandrashekhar, and Vijay Kumar Jha. "Decision tree and genetic algorithm based intrusion detection system." In Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017), pp. 141-152. Springer, Singapore, 2019.

[12] Alom, Md Zahangir, Venkata Ramesh Bontupalli, and Tarek M. Taha. "Intrusion detection using deep belief networks." In 2015 National Aerospace and Electronics Conference (NAECON), pp. 339-344. IEEE, 2015.

[13] Mohammadi, Shahriar, and Fatemeh Amiri. "An efficient hybrid self-learning intrusion detection system based on neural networks." International Journal of Computational Intelligence and Applications 18, no. 01 (2019): 1950001.

[14] Musavi, Seyyedeh Atefeh, and Mahmoud Reza Hashemi. "HPCgnature: a hardware-based

application-level intrusion detection system." IET Information Security 13, no. 1 (2018): 19-26.

[15] Kumar, Vikash, Ditipriya Sinha, Ayan Kumar Das, Subhash Chandra Pandey, and Radha Tamal Goswami. "An integrated rule-based intrusion detection system: Analysis on UNSW-NB15 data set and the real-time online dataset." Cluster Computing (2019): 1-22.

[16] Anshul Chaturvedi and Prof. Vineet Richharia "A Novel Method for Intrusion Detection Based on SARSA and Radial Bias Feed Forward Network (RBFFN)", in international journal of computers & technology vol 7, no 3.

[17] Jain, Upendra "An Efficient intrusion detection based on Decision Tree Classifier using feature Reduction", International Journal of Scientific and Research Publications, Vol. 2, Jan. 2012.

[18] E. Blanzieri and A. Bryl "A survey of learning-based techniques of email spam filtering" Artif. Intell. Rev., vol. 29, no. 1, pp. 63–92, 2008.

[19] D. Sculley and G. Cormack "Filtering email spam in the presence of noisy user feedback" in Proc. 5th Email Anti-Spam Conf., 2008, pp. 1– 10.

[20] HengjieLi, Jiankun Wang "Intrusion Detection System by Integrating PCNN and Online Robust SVM" IFIP International Conference on Network and Parallel Computing, 2007.pp 250-255.