International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 267-272

Investigation of Machine Learning Techniques for Intrusion Detection

Mrs.G.Janani @ Pandeeswari
Assistant Professor,Department Of Computer Science And Engineering,RVSETGI,Dindigul

Abstract
The Internet of Things (IoT) consists of many devices that are connected via the internet and can interact with each other. IoT applications change our work and lives by saving time and resources. IoT also offers unlimited benefits and provides easy access to the information you need. While it is the fastest growing technology, it introduces security concerns. The Internet of Things is vulnerable to various types of cyber attacks. Most of the current research work is based on network penetration detection. The recent development of machine learning techniques has played a key role in effectively detecting intruders. Intrusion Detection System is a system that monitors the network and detects suspicious activity. This article contributes to a comprehensive study of IDS using machine learning
Keywords*: Intrusion Detection System · Machine learning · Classification.*

Date of Submission: 25-05-2022 Date of acceptance: 05-06-2022

I. Introduction

Security breaches are increasing significantly in cyberspace, impacting cybersecurity. The number of network attacks is increasing every day. Researchers have been proposing methods to detect and prevent such security breaches for more than four decades. When designing network protocols, it is important to ensure network reliability. Traditional cybersecurity prevention tools such as antivirus software and firewalls are not enough to prevent or detect intruders.]. Intrusion detection is a dynamic process to monitor network progress, analyze and intimidate many malicious events from unauthorized access [4]. This can be achieved by gathering information from various system and network sources and then analyzing information about potential security issues. To resolve this issue, the network must continuously monitor for problems with the delegated software. Intrusion Detection Systems (IDS) are an important part of computer security systems [33]. The job of an IDS is to detect malicious activity in a computer system and thereby enable a rapid response to attacks. Anderson [4] notes the following behavior identified by IDS in "Computer Security Threat Monitoring Systems". Exceptional use, unusual frequency of use, unusual amount of reference data, and unusual references to program or data.

According to IDC CyberGuidelines for Security Expenditure 2019, global security spending could exceed US$100 billion in 2019 [13]. According to Kaspersky Security Lab, the number of denials of service has increased since 2013 [43], but attackers are already focusing on more advanced attacks [25]. 2 damn questions. On the other hand, more and more sophisticated attack techniques were discovered from time to time. As a result, traditional methods based on signatures and expert rules are no longer adequate [34]. IDS based on machine learning techniques have been developed to address security concerns [33]. These techniques benefit from recent developments in the machine learning research community and the wide range of data collected in cybersecurity research. The drawback of existing methods based on machine learning techniques is that they are trained on a single set of data, making new types of attacks more difficult. These techniques require large amounts of data that take time to collect. On the other hand, large machine learning algorithms in IDS have to be trained from scratch when new data arrives, which requires high computing power. Although researchers have designed a comprehensive machine learning system for IDS, there is still room for improvement. The main purpose of this article is to present a study of machine learning techniques and intrusion detection systems. Required concepts related to IDS are examined. First, the concept of IDS is presented, in the definition of which its types and meanings are discussed. Then review machine learning techniques. This document provides an overview of electronic machine learning techniques for IDS.

II. Related Work

This section provides information on three important topics necessary for understanding the following sections of the article: Intrusion Detection Systems, Literature Review

2.1 Intrusion Detection System

As Internet use increases, so does the potential for cyber attacks. In many cases, these attacks are new and require intelligent systems to detect them. Intrusion detection is the process of monitoring and analyzing network or computer traffic for signs of attack
penetration [2].

The main objectives of IDS can be summarized as follows:

1) Host and network monitoring,
2) analysis of computer network behavior,
3) generate signal,
4) Respond to suspicious behavior.

IDSs are divided into two main types: abuse-based IDS and anomaly-based IDS. Abuse-based detection is compared to existing attacks on current traffic, and alerts are triggered if there is a match. On the other hand, anomaly-based detection monitors the network for any deviation from normal behavior and reports it as an anomaly or abnormal. Anomaly-based detection is important to identify zero-based attacks [24].
problem ID:

1. High False Alert Rate: When an alert is triggered by a non-threatening violation and a very serious violation is not detected.
2. New attacks are difficult to detect, which has drawn attention to the use of modern machine learning techniques to detect intruders.

III. Literature Review

There are two main approaches to using machine learning for IDS: the controlled learning approach, in which the model tries to distinguish between normal traffic and malicious traffic, and the unsupervised learning approach, which is primarily based on the anomaly detection method that the model tries to distinguish between normal. traffic and other traffic.

3.1. Approach to supervised learning

Most of the ready-to-use controlled learning approaches were used in the IDS study [22]. In short, the main task of controlled classifiers is to predict normal or malicious attacks on network flows. To identify malicious flows, a set of data streams is first defined which must be trained first.

Traditional approach to machine learning

Traditional machine learning approaches such as SVM, logistic regression, or decision trees have been used in IDS for a long time. Authors in [28, 26, 3, 42] propose the use of solution trees for classification problems. A decision tree is a classification algorithm that aims to create a consistent if-else rule by minimizing the loss function when breaking up groups. The most commonly used loss function when constructing a decision tree is Gini [3]. While authors [28, 26, 3] construct decision trees only on a given data set, [42] combine decision tree techniques with genetic algorithms [30] to generate features to achieve better predictive efficiency.

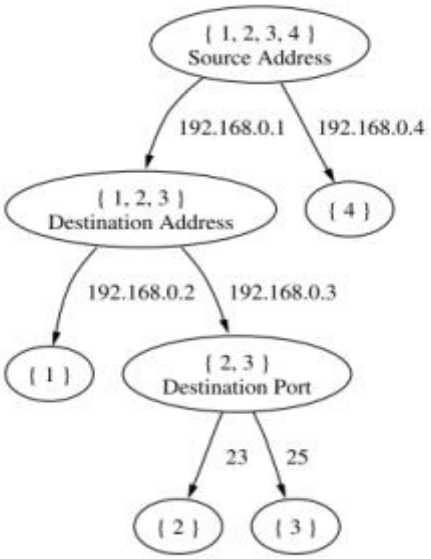


Fig. 1. A decision tree built by Kruegel et al. [26]

Another popular traditional classification algorithm that is widely used in IDS is the Support Vector Machine (SVM). The basic principle of SVM is to find an optimized hyperplane that can classify between two classes. In practice, the most difficult task when building an SVM model is finding a good kernel. In the early 2000s, a combination of several techniques was proposed. Through the development of ensemble training techniques, random forest was chosen as the most widely used technique [36]. The idea of each forest is to build several decision trees and then combine the predictions of each tree. Random forest has been shown to outperform other classification techniques in well-defined datasets [16].The random forest algorithm is shown in Figure 2.

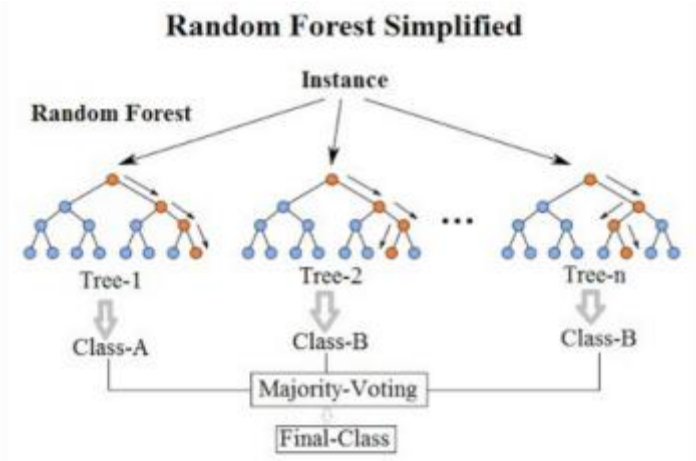
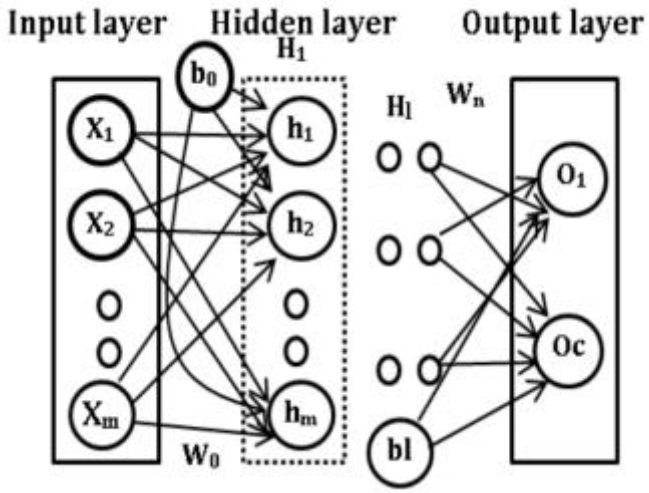


Fig. 2. Random Forest Deep Learning Based Approaches

In recent years, research on deep learning has yielded several important breakthroughs [16-18, 15]. Not surprisingly, researchers are harnessing the power of deep learning methods in IDS. However, redirection neural networks are mainly used in IDS. A multilayer neural network with redirection is shown in Figure 3.



The authors in [19] propose to use multilayer neural networks in a detour rather than traditional methods such as logistic regression or SVM in IDS. The author [45] considers not only the function generated by the system log, but also the use of word insertion techniques to learn from system calls. Apart from simple multilayer neural networks with switching, several research studies have used more complex networks such as ConvNet or Recurrent Neural Networks [46].

3.2 Approaches to detect anomalies

In contrast to surveillance-based approaches, the anomaly detection approach does not use a labeled data set but instead tries to investigate what is benign flow and detects abnormal tissue flow [14]. Therefore, the main assumption of the anomaly detection algorithm is that the benign flow represents most of the classes in the data. Anomaly detection algorithms can be divided into two subgroups: statistical-based algorithms and machine

learning-based algorithms. Statistical algorithms rely on statistical data attributes such as z-scores. Algorithms for detecting anomalies in machine learning rely on machine learning techniques to examine the normal data model and then determine whether a new instance belongs to the normal data class or not. Eskin et al. [20] defined an SVM class (OCSVM) based on the idea of SVM in classification, but instead of separating the two classes, the OCSVM algorithm tries to separate the data from its origin. Liu et al. [29] used the idea of an arbitrary forest to construct an isolated forest algorithm that determines the resulting deviation from the tree depth required to classify a single data point. It is difficult to classify normal points in a data set because they are very similar to other points, whereas classifying anomalous points is relatively easy. A large number of studies on deviation detection can be found in [1]. There are several studies to detect abnormalities, especially for IDS, such as [10] and [6]. In general, anomaly detection algorithms can have better generalization than controlled algorithms because they do not require labeled data and can be adapted to the dynamic nature of attacks. However, the main concern of the anomaly detection algorithm is the long run time, which makes it impractical for real-time systems [44].

IV. Machine learning and IDS

The authors in [24] examined the performance of ten classification algorithms using the NSL-KDD dataset, using a different approach to feature selection than [12]. The feature selection approach is based on the implementation of attribute evaluators and filters. The author applies the algorithm Naive Bayes, Bayes-Net, Logistics, Random Tree, Random Forest, J48, Bagging, OneR, PART, ZERO. The most effective classification algorithm is random forest with an accuracy of 99.9% and a low percentage of false positives of 0.001. The second best classifier is Pocket with 99.8% accuracy and this performance is reported by the PART algorithm.

Belavagi, M.C., & Muniyal, B.in [25] used the NSL-KDD dataset and applied a retention test approach without using a feature selection approach. The following four classification algorithms were tested: random forest, SVM, logistic regression, and mixed Gaussian model. Random Forest turned out to be the best algorithm with 99% accuracy. The second best classifier is logistic regression, with a reported accuracy of 84%.

Farnaaz, N., and West Java, MA (2016). et al. [12] describes the problem of classifying offenders using the NSL-KDD data set. First, the dataset is pre-processed to fix the missing data problem and categorize the numeric attributes. The data sets were then grouped into four data sets and divided into training data and test data. The data is then sent to each forest classifier and the classification and accuracy and FPR are calculated. A feature selection approach using a symmetrical uncertainty measure is also applied. The author reports a slight performance improvement after implementing the feature selection. Reported results are compared with C4.5, but each forest better than C4.5 [12]. After selecting a function, the average reported accuracy is 99.67 and the percentage of false positives is 0.005.

The authors in [26] experimented with artificial neural networks (ANN) and SVM representations on the NSL-KDD sample dataset. The sample represents 20% of the entire data set. Two methods were used to select features: based on correlation and based on the chi-square method. The first leads to a choice of 17 functions and the second to 35 functions. After selecting a function, the data is sent to the ANN and SVM classifiers. The results with the selection of correlation-based functions and ANN showed the best performance with an accuracy of 94.0%.

DoS, U2R and R2L.

The CFSSubSet Eval and Best First feature selection algorithms were used in combination with four methods: uncontrolled, k-medium, and three controlled methods, SVM, Naive Bayes, and Random Forest. Three controlled methods outperform unsupervised techniques and are based on eight best traits. The best reported accuracy is Random Forest with 99.0% accuracy.

KDD'99 was also used by the authors in [28], where the Ant Colony algorithm was first applied to select a suitable representative set from an original data set of 550 samples (note). The author then applies a new feature reduction method, called the phase reduction method (GFR), to reduce the dimensions of the feature space to 19 features. The reduced features are then combined with SVM for classification. The reported accuracy is 98.67% before the feature is selected and there is no significant increase in accuracy after the feature is selected and the result is 98.62%.

The authors in [26] experimented with artificial neural network (ANN) and SVM representations on the sample NSL-KDD dataset. The sample represents 20% of the entire dataset. Two methods were used to select features: based on correlation and based on the chi-square method. The first leads to a choice of 17 features and the second leads to up to 35 features. After selecting a function, the data is sent to the ANN and SVM classifiers. The results with the selection of correlation-based functions and ANN showed the best performance with an accuracy of 94.0%.

The KDD'99 data set was used by the authors in [27] with classes: normal, prob,

DoS, U2R and R2L. The CFSSubSet Eval and Best First feature selection algorithms were used in combination with four methods: The CICIDS2017 data set used by the authors in [30, 31] applies a sample-

based, power-reduction, and enhancement-based approach to creating IDS. Datasets have a big problem with dataset imbalances. In doing so, they pre-processed the data using the Synthetic Minority Oversampling Technique (SMOT) to handle a small number of cases across multiple classes. They then applied a reduction approach based on ensemble feature selection (EFS) and principal component analysis (PCA) to a reduced feature space of 25 features. The results are based on the Adaboost algorithm and evaluated according to the retention method. Results are reported with an accuracy of 81.83%.

Recently [32] the authors discussed the problem of classifying attacks using random forests and ANN techniques based on the CICIDS2017 dataset. They implemented a package called Boruta for feature selection and it returns the top 10 features. The feature set is then presented to the classifier. They reported an average accuracy of 96% using ANN and 96.4% using any forest.

Another dataset that attracted attention in this domain is the UNSW-NB15. In the work of [33], the authors targeted this data set with an approach that uses k-means clustering, CFS feature selection, and four different techniques SVM, RF, J48, and Zero. The proposed approach has been effective in improving the performance of the majority of classifiers. The best-reported accuracy is using J48 with an accuracy of 96.7% and using 10-fold cross-validation.

The authors of [34] targeted also the network intrusion detection problem on the UNSW-NB15.

V. Conclusion

Cybercrime is on the upward push, manner to the latest surge in internet content. The use of intrusion detection systems (IDS) is the initial step in detecting and reporting such assaults. The detection of anomalies is reliant on detecting unique assaults, that is a difficult task. This has piqued the interest of students everywhere in the global who need to research more about this subject and, specifically, a way to use supervised gaining knowledge of algorithms for intrusion detection to move over IDS class, supervised getting to know techniques, and cyber protection attacks great in this check.Then, the use of four famous datasets: KDD'99, NSL-KDD, CICIDS2017, and UNSW-NB15, we summarized related efforts in this difficulty. Supervised studying algorithms' classification performance is right and promising, in accordance toa observe of four records sets: Furthermore, characteristic selection is vital and, in masses of conditions, required for general overall performance development. Data imbalance can be a scenario, and sampling techniques can help remedy the problem. Finally, for unique overall performance, big intrusion detection records units necessitate a deep getting to know .

References

[1]. Graham, J., Olson, R., & Howard, R. (Eds.). (2011). Cyber security essentials. CRC Press.
[2]. Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.
[3]. Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. applied sciences, 9(20),4396
[4]. Hamid, Y., Sugumaran, M., & Balasaraswathi, V. R. (2016). Ids using machine learning-current state of the art and future directions. Current Journal of Applied Science and Technology, 1-22.
[5]. Masdari, M., & Khezri, H. (2020). A survey and taxonomy of the fuzzy signature-based intrusion detection systems. Applied Soft Computing, 106301.
[6]. Milenkoski, A., Vieira, M., Kounev, S., Avritzer, A., & Payne, B. D. (2015). Evaluating computer intrusion detection systems: A survey of common practices. ACM Computing Surveys (CSUR), 48(1), 1-41.
[7]. Conrad, E., Misenar, S., & Feldman, J. (2012). CISSP study guide. Newnes
[8]. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
[9]. Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record, 31(1), 76-77.
[10]. https://en.wikipedia.org/wiki/, accessed 1/5/2021
[11]. Sahasrabuddhe, A., Naikade, S., Ramaswamy, A., Sadliwala, B., & Futane, P. (2017). Survey on intrusion detection system using data mining techniques. Int Res J Eng Technol, 4(5), 1780-4.
[12]. Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.
[13]. Rust, J. (1997). Using randomization to break the curse of dimensionality. Econometrica: Journal of the Econometric Society, 487-516.
[14]. https://www.unb.ca/cic/datasets/index.html, accessed 1-6-2021
[15]. www.kaggle.com, accessed 1-6-2021
[16]. Uma, M., & Padmavathi, G. (2013). A Survey on Various Cyber Attacks and their Classification. IJ Network Security, 15(5), 390-396.
[17]. Li, X., Smith, J. D., Dinh, T. N., & Thai, M. T. (2016, October). Privacy issues in light of reconnaissance attacks with incomplete information. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 311-318). IEEE.
[18]. Hussain, A., Heidemann, J., & Papadopoulos, C. (2003, August). A framework for classifying denial of service attacks. In Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications (pp. 99-110).
[19]. Forcht, K. A., Kieschnick, E., Thomas, D. S., & Shorter, J. D. (2007). Identity Theft: The Newest Digital Attack. Issues in Information Systems, 8(2,297-302).
[20]. https://en.wikipedia.org/wiki/Cyber_spying#Examples, accessed 20-5-2021

[21]. Bhardwaj, A., & Goundar, S. (2020). Keyloggers: silent cyber security weapons. Network Security, 2020(2), 14-19.
[22]. Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. Journal of management information systems, 28(2), 203-236.
[23]. Simmons, C., Ellis, C., Shiva, S., Dasgupta, D., & Wu, Q. (2014, June). AVOIDIT: A cyber attack taxonomy. In 9th Annual Symposium on Information Assurance (ASIA'14) (pp. 2-12).
[24]. Malhotra, H., & Sharma, P. (2019). Intrusion Detection using Machine Learning and Feature Selection. International Journal of Computer Network & Information Security, 11(4).
[25]. Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion de tection. Procedia Computer Science, 89, 117-123.
[26]. Taher, K. A., Jisan, B. M. Y., & Rahman, M. M. (2019, January). Network intrusion detection using supervised machine learning technique with feature selection. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 643 -646). IEEE.
[27]. El Mourabit, Y., Bouirden, A., Toumanari, A., & Moussaid, N. E. (2015). Intrusion detection techniques in wireless sensor network using data mining algorithms: comparative evaluation based on attacks detection. International Journal of Advanced Computer Science and Applications, 6(9), 164-172.
[28]. Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert systems with applications, 39(1), 424-430.
[29]. Shah, B., & Trivedi, B. H. (2015, February). Reducing features of KDD CUP 1999 dataset for anomaly detection using back propagation neural network. In 2015 Fifth International Conference on Advanced Computing & Communication Technologies (pp. 247-251). IEEE.
[30]. Yulianto, A., Sukarno, P., & Suwastika, N. A. (2019, March). Improving Adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012018). IOP Publishing.
[31]. Abdulhammed, R., Faezipour, M., Musafer, H., & Abuzneid, A. (2019, June). Efficient network intrusion detection using pc a-based dimensionality reduction of features. In 2019 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE.
[32]. Pelletier, Z., & Abualkibash, M. (2020). Evaluating the CIC IDS-2017 Dataset Using Machine Learning Methods and Creating Multiple Predictive Models in the Statistical Computing Language R. Science, 5(2), 187-191.
[33]. Hammad, M., El-medany, W., & Ismail, Y. (2020, December). Intrusion Detection System using Feature Selection With Clustering and Classification Machine Learning Algorithms on the UNSW-NB15 dataset. In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) (pp. 1-6). IEEE.
[34]. Faker, O., & Dogdu, E. (2019, April). Intrusion detection using big data and deep learning techniques. In Proceedings of the 2019 ACM Southeast Conference (pp. 86-93).