Examination of Port Scan Detection Algorithms Utilizing Deep Learning and Support Vector Machines

Dr. V Venkata Ramana[1,] A Ram Praksh Reddy[2], B Swetha[3] Dr. V. Lokeswara Reddy[4]

[1,4] Professor, [2,3] Asst. Professor,
Department of CSE, K.S.R.M College of Engineering(A), Kadapa

Abstract—

Recent technological developments in computer and communication have resulted in vast advancements over their forerunners. The use of cutting-edge technology has numerous beneficial outcomes for individuals, organizations, and states, but it also has certain drawbacks. Knowledge accessibility, data privacy, and data storage security are just a few of the concerns that have been brought up as causes for concern. These elements have elevated cyber terrorism to the forefront of contemporary security concerns. Acts of cyber terrorism have already caused enormous disruptions to people's lives and institutions, and they are now capable of being committed by a wide variety of parties, including criminal organizations, professionals, and online activists. For this reason, Intrusion Detection Systems (IDS) have been developed to keep harmful programs out of computer networks. This study, which used the most recent CICIDS2017 dataset, successfully identified port scan attempts using deep learning (accuracy: 97.80 percent) and support vector machine (SVM) methods (accuracy: 69.79 percent).

INTRODUCTION

The number of computer crimes has been rising steadily. They are not limited to harmless activities like guessing a system's login credentials but instead pose a far greater threat. The goal of information security is to prevent loss, misuse, alteration, or disclosure of sensitive data. Information security, computer security, and data insurance are all common synonyms. These

Each subfield contributes to the overall mission of ensuring information accessibility, security, and privacy. Discovery has been shown to be the starting point for any assault [1]. In this phase, reconnaissance efforts are done to learn more about the system. For an attacker, discovering a system's list of open ports is a game-changer. This is why there are so many antivirus and IDS programs devoted to port scanning [2]. In this study, we built IDS models to identify port scan attempts using deep learning and SVM machine learning methods. There was some contrasting between the models shown. We divided the rest of the paper into the following sections: In Section 2, we provided our literature review. The materials and procedures utilized were described in Section 3. In Section 4, we presented experimental findings from our work on categorization algorithms and performance metrics. The last section summed everything up and discussed what's next.

LITERATURE REVIEW

Information security concepts consist of human, period, methodology, knowledge, system and technology as is shown in Figure 1. Confidentiality, integrity, and accessibility have to be provided by a secure system. First, the confidentiality of the information means allowing access only to the person who needs to access that information. Second, the integrity of the information is ensuring that the information is protected without distortion and the original structure is intact. Finally, the accessibility of information is the ability to access and use information at the desired time.

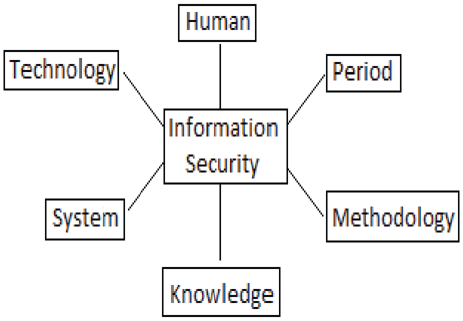


Fig. 1. Information security concepts [3].

Stanford et al. [4] indicate that surprisingly little research has been done on the topic of detecting port scans. Robertson et al. [5] used a threshold technique to identify the unsuccessful connections. To locate the intruder using the NSL-KDD dataset, Ibrahim and Outdone [6] used Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). Mustafa and Slay [7] demonstrated the contrasting results of using the KDD99 and UNSW-NB15 datasets to investigate network behaviour. Using the KDD99 dataset, Laying et al. [8] identified and categorized harmful patterns in network traffic. Alanson and Lumet [9] used Naive Bays and Principal Component Analysis (PCA) on the KDD99 dataset. Chitin and Rabbinic for IDS also employed PCA, SVM, and KDD99 [10]. For their IDS model, Aljawarneh et al. [11] used the NSLKDD dataset for analysis and experimentation. There is a consistent usage of the KDD99 dataset for IDS in the literature [6]–[10]. KDD99 was created in 1999 and has 41 individual features. This is why KDD99 is out of date; it doesn't cover the most recent developments in cyberattacks like zero-day vulnerabilities. As a result, we conducted our research using the fresh CICIDS2017 dataset [12]. The number of research using data from CICIDS2017 is small but growing. This article has already covered some of them. Using the CICIDS2017 dataset, D. Aksum et al. [13] demonstrated the effectiveness of several machine learning techniques in identifying Dodos attacks. They did not utilize the whole dataset but instead used only 26.167 Dodos and 26.805 harmless samples. In addition, the Fisher score feature selection technique was utilized to zero in on the most useful characteristics. As a consequence, their prior SVM models achieved a high level of accuracy. However, they intended to use a deep learning system as part of its detection features in order to identify Dodo assaults. Distributed research was suggested by N. Mari et al. [14] to identify anomalous behavior in a global network. Resend et al. [15] employed evolutionary algorithms to find malicious activity in the CICIDS2017 dataset.

MATERIAL AND METHODS

The CICIDS2017 dataset and deep learning and SVM algorithms are explained respectively in this section.

A. CICIDS2017 Dataset The CICIDS2017 dataset is used in our study. The dataset is developed by the Canadian Institute for Cyber Security and includes various common attack types. In this study, we focused on port scan attempts. There are 286467 records consisting 127537 benign and 158930 port scan attempts and each record has 85 features such as source IP, source port, destination port, flow duration, total fwd packets, total backward packets etc. A part of the records is as shown in Table I.

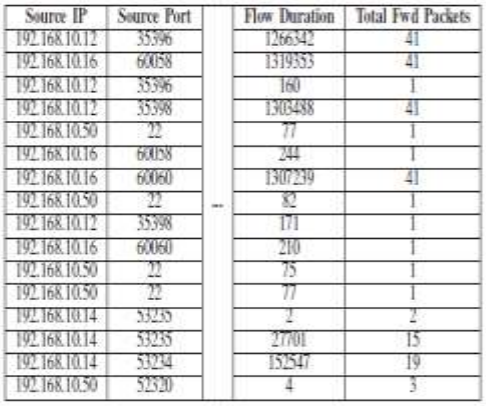
When creating the dataset, Attack-Network and Victim- Network, completely were separated two networks, were designed and implemented by Sharafaldin H. et al [12]. They collected data from July 3, 2017, to July 7, 2017, for the dataset.

B. SVM

Statistical learning and convex optimization, based on the principle of structural risk minimization, form the basis of Support Vector Machine (SVM) algorithms. Vapid et al developed SVM as a solution to different problems [16]. For example, it can be used in many different areas such as learning, pattern recognition, regression, classification, and analysis.

TABLE I

A SAMPLE SET OF RECORDS FROM DATASET [12]



Source IP	Source Port		Flow Duration	Total Fwd Packets
192.168.10.12	35396	...	1266342	41
192.168.10.16	60058		1319353	41
192.168.10.12	35396		160	1
192.168.10.12	35398		1303488	41
192.168.10.50	22		77	1
192.168.10.16	60058		244	1
192.168.10.16	60060		1307239	41
192.168.10.50	22		82	1
192.168.10.12	35398		171	1
192.168.10.16	60060		210	1
192.168.10.50	22		75	1
192.168.10.50	22		77	1
192.168.10.14	53235		2	2
192.168.10.14	53235		27701	15
192.168.10.14	53234		152547	19
192.168.10.50	52320		4	3

SVM is a supervised learning method because it uses tagged data in a dataset as an input. The number of output classes changes depending on the dataset. For example, two classes of output data are generated when a dataset of two classes is given as the input. Therefore, the samples given as the input are categorized according to these classes. During the training process, a model is created according to the input dataset and classification is performed by using the model.

C. Deep Learning

Deep Learning algorithms allow to extract features automatically from a given dataset and they consist of a sequential layer architecture. Applying non-linear transformation functions to the sequential layer structure constitute the basis of deep learning algorithms. Increasing the number of layers will increase the complexity of nonlinear transformations to be constructed. Deep learning algorithms learn the abstract hidden properties of the data obtained in the last layer from its abstract representations acquired at multiple levels. Therefore, the abstract properties of the final layer’s output are obtained by introducing the data into a high-level non-linear function.

D. Methodology

The SVM and deep learning algorithms were used to detect port scan attempts based on the CICIDS2017 dataset. The flowchart of the proposed method was presented in figure 2. First of all, 286.467 records which consist of 158.930 port scan attempts and 127.537 benign behaviors are taken from the dataset and then these records were normalized. After normalization samples were split into two as a 67% training data and 33% testing data. In addition, the SVM and deep learning IDS models were created based on the training data. Finally, the models were tested with test data and the performance of models was calculated comparatively. In addition, the deep learning IDS model consist of 7 hidden layers and each layer include the different number of neurons such as 100, 150, 70, 40 and 6 respectively. The rely was selected and used as an activation function in the model. Depending on the number of neurons and hidden layer model performances were changed. In this paper, we selected optimum numbers based on the model’s accuracy. On the other hand, we did not apply any feature selection algorithm for SVM and we used all features. As a future work, we are going to use different artificial intelligence approaches to define select this optimum values.

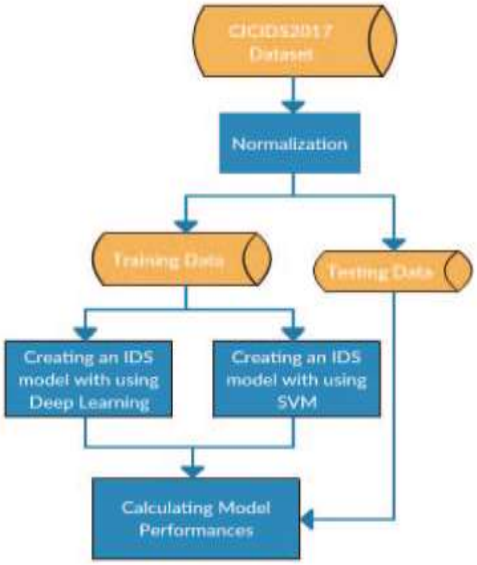

CICIDS2017
Dataset
Normalization
Training Data
Testing Data
Creating an IDS model with using Deep Learning
Creating an IDS model with using SVM
Calculating Model Performances

Fig. 2. Flowchart of the method.

As is shown in figure 2, main steps of the algorithm are presented in below.

1) Normalize the dataset.

2) Split the normalized dataset into two as training and testing.

3) Create IDS models with using SVM and deep learning algorithms.

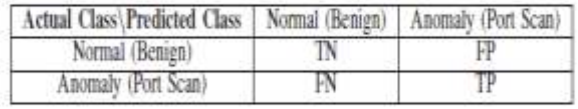
4) Evaluate the models' performances. In normalization, nonnumeric label features were converted into numeric forms. In addition, unrelated features such as Timestamp and some samples that have Nan, infinity and empty values were removed. Furthermore, we rescaled all observed values of features to have a length of 1. As a second step, the normalized dataset was split into 67% training and 33% testing. In the third step, the IDS models were trained and generated to detect port scan attempts by using the training data. Consequently, the performances of the models were calculated. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) statistics (Table II) are used for evaluation of model performances.

Table II can be explained in below items.

- TN: Actual Benign is classified as Benign.
- FP: Actual Benign is classified as Port Scan.
- FN: Actual Port Scan is classified as Benign.

TABLE II

CONFUSION MATRIX

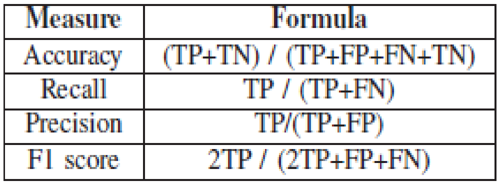


Actual Class\Predicted Class	Normal (Benign)	Anomaly (Port Scan)
Normal (Benign)	TN	FP
Anomaly (Port Scan)	FN	TP

• TP: Actual Port Scan is classified as Port Scan. Accuracy, recall, precision and f1 score performance metrics are calculated using the statistics of the confusion matrix (Table III).

TABLE III

PERFORMANCE METRICS [17]



Measure	Formula
Accuracy	(TP+TN) / (TP+FP+FN+TN)
Recall	TP / (TP+FN)
Precision	TP/(TP+FP)
F1 score	2TP / (2TP+FP+FN)

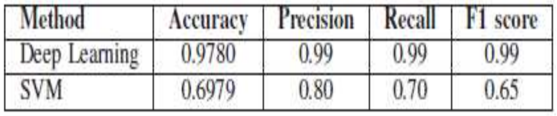
The ratio of correctly predicted observations is accuracy, while precision means a ratio of correct positive observations. The recall is a proportion of correctly predicted positive events. F1 score signifies the weighted average of precision and recall.

EXPERIMENTAL RESULTS

The personal computer which has Intel(R) Core(TM) i7- 5700HQ CPU @2.70 GHz, 16 GB Ram capacity was used for experiments. We used the CPU; however, we are considering applying GPU as a future work. 286.096 records, which were taken from the normalized dataset, were divided into two sets with 67% training and 33% testing ratios such as 191684 samples for training and 94412 samples for testing. The deep learning model was trained in 30 Epochs and performance measurement of the SVM and deep learning models presented in Table IV.

TABLE IV

PERFORMANCE METRICS OF USED CLASSIFICATION TECHNIQUES BASED
ON CICIDS2017 DATASET.



Method	Accuracy	Precision	Recall	F1 score
Deep Learning	0.9780	0.99	0.99	0.99
SVM	0.6979	0.80	0.70	0.65

Table IV shows the accuracy, recall, precision and F1 score rates of the IDS models which were developed by using deep learning and SVM. Deep learning achieved a higher success than SVM.

CONCLUSION AND FUTURE WORKS

In this research, we give a comparison of support vector machine and deep learning algorithms' performance metrics using the most recent CICIDS2017 dataset. The findings demonstrate that the deep learning algorithm outperformed the SVM by a wide margin. Based on this information, we want to leverage machine learning and deep learning techniques, Apache Hadoop and Spark technology, and various attack types beyond port scan efforts.

REFERENCES

[1] K. Graves, C.E.H.T. : The Official Study Guide for the Certified Ethical Hacker Examination, Version 312-50. Published by John Wiley & Sons, 2007.

"Port scanning techniques and the defense against them," by R. Christopher (2001) at the SANS Institute.
For more information on this topic, see: [3] M. Bayar, R. Das, and I. Karado gan, "Bilge g uvenli gi sistemlerinde kullanlan arac,larn incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231-239.

According to [4] "Practical automated detection of stealthy portscans," written by S. Staniford, J. A. Hoagland, and J. M. McAlerney and published in Journal of Computer Security, vol. 10, no. 1-2, pp. 105-136, 2002.

[5] "Surveillance detection in high bandwidth environments," by S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, published in DARPA Information Survivability Conference and Exposition, 2003. Proceedings, volume 1. IEEE, 2003, pages 130-138.

Reference: [6] K. Ibrahimi and M. Ouaddane, "Management of intrusion detection systems based-kdd99: Analysis with lda and pca," in Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

"The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," by N. Moustafa and J. Slay. IEEE, 2015. Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on.

Reference: [8] "Detection and classification of malicious patterns in network traffic using benford's law," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864-872. Authors: L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang.

[9] S. M. Almansob and S. S. Lomte, "Addressing challenges for intrusion detection system using naive bayes and pca algorithm," 2017 2nd International Conference on Convergence in Technology (I2CT), IEEE, pp. 565-568.

As cited in [10] "Combined analysis of support vector machine and principle component analysis for ids," published in IEEEInternational Conference on Communication and Electronics Systems, 2016, pages 1-5. Authors: M. C. Raja and M. M. A. Rabbani.