Port Scan Detection using AI

Yawar Rasool Mir
Department of Computer Science & Engineering
Desh Bhagat University
Punjab, India

Navneet Kaur Sandhu
Department of Computer Science & Engineering
Desh Bhagat University
Punjab, India

Abstract: Contrasted with past, developments in computer and communication technologies have provided extensive and advanced changes. The use of new advances give incredible advantages to people, organizations, and governments, in any case, it causes a few issues against them. For example, the protection of significant data, security of put away information stages, accessibility of learning and so on. Contingent upon these issues, digital fear based oppression is a standout amongst the most significant issues in this world.

Cyber fear, which made a ton of issues people and establishments, has achieved a dimension that could undermine open and nation security by different gatherings, for example, criminal associations, proficient people and digital activists. Hence, Intrusion Detection Systems (IDS) have been created to maintain a strategic distance from digital assaults. In this examination, Artificial Neural Network, Random Forest (RF) and Support vector machine (SVM), calculations were utilized to recognize port scan attempts dependent on the new CICIDS2017 dataset and 98.87%, 99.20%, 72.19% precision rates were achieved respectively.

***IndexTerms* — IDS, Cyber Terror, ANN, SVM, RF, CICIDS2017**

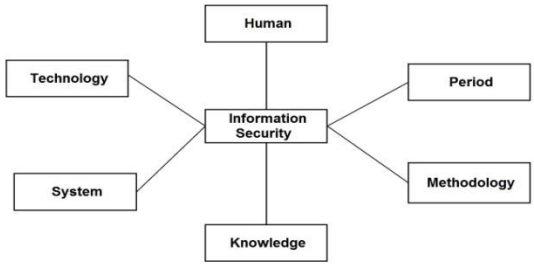
I. INTRODUCTION

Computer crime's keeps on expanding throughout the years. They are not just limited to unimportant acts, for example, evaluating the login qualifications of a framework yet additionally they are substantially more hazardous. Data security is the way towards shielding data from unapproved get to, utilization, revelation, devastation, alteration or harm. The terms" Information security" "PC security" and "information protection" are regularly used interchangeably. These areas are related to each other and have common goals to provide availability, confidentiality, and integrity of information. Studies show that the first step of an attack is discovery [1]. Reconnaissance is made in order to get information about the system in this stage. Finding a list of open ports in a system provides very critical information for an attacker. For this reason, there are a lot of tools to identify open ports [2] such as antivirus and IDS.

In this work, deep learning, RF and SVM machine learning algorithms were applied to create IDS models to identify port scan attempts. The models were exhibited relatively arranged different pieces of the paper as: a literature review was presented an explanation of used material and methods. Experimental results of the classification algorithms and estimations were presented in Section 4. Segment 5 gave end and future works.

II. LITERATURE REVIEW:

Data security concepts consist of human, period, strategy, information, framework and innovation as is shown in Figure 1. Privacy, integrity, and availability must be give secure system. In the first place, the privacy of the data implies permitting access just to the individual who needs to get to that data. Second, the respectability of the data is guaranteeing that the data is ensured without bending and the first structure is unblemished. At last, the availability of data is the capacity to access and utilize data at the ideal time.



Information security concepts

As is connoted by Stanford et al, there has been amazingly restricted work on the issue of identifying port scan [4].Robertson et al. utilized a limit technique to recognize the fizzled association attempts [5].Linear Discriminate Analysis (LDA) and Principal Component Analysis (PCA) were applied by Ibrahimi and Ouaddane to identify the intrusion with NSL-KDD dataset [6]. Near outcomes of KDD99 and UNSW-NB15 data sets analyzing network behaviors were showed by Mustafa and Slay [7]. Liuying et al. detected and characterized noxious examples in system traffic dependent on the KDD99 dataset [8]. Guileless Bayes and Principal Component Analysis (PCA) were utilized with the KDD99 dataset by Almansob and Lomte [9].Similarly, PCA, SVM, and KDD99 were utilized Chithik and Rabbani for IDS [10]. In Aljawarneh et al's. paper, their investigation and analyses were created dependent on the NSL-KDD dataset for their IDS model [11]

Literature studies show that KDD99 dataset is constantly used for IDS [6]–[10]. There are 41 includes in KDD99 and it was created in 1999. Therefore, KDD99 is old and does not give any data about exceptional new assault types, for example, multi day abuses and so forth. In this way we utilized a forward-thinking and new CICIDS2017 dataset [12] in our examination. There are distinctive however constrained examinations dependent on the CI-CIDS2017 dataset. Some of them were talked about here. D.Aksu et al. demonstrated exhibitions of different AI calculations identifying DDoS assaults dependent on the CICIDS2017 dataset in their past work [13]. They didn't matter all dataset and utilized restricted information 26.167 DDoS and 26.805 kind examples from the dataset in their investigation. Also, they utilized the Fisher score highlight choice calculation to choose the best highlights. In this way, their past SVM models achieved a high exactness result. Be that as it may, they were intending to apply profound learning calculation as an element work to identify DDoS assaults. N. Marir et al. proposed a conveyed examination to find unusual action in an large scale network [14]. In another examination, Resende et al. used genetic algorithms to recognize interruptions on the CICIDS2017 dataset [15]

III. MATERIAL AND METHOD:

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

A. CICIDS2017 Dataset

The CICIDS2017 dataset is utilized in our examination. The dataset is created by the Canadian Institute for Cyber Security and incorporates different basic assault types. In this examination, we concentrated on port scan attempts. There are 286467 records comprising 127537 amiable and 158930 port scan attempts and each record has 79 highlights, for example, source IP, source port, goal port, stream length, complete fwd bundles, all out in reverse parcels and so on. A piece of the records is as appeared Table I.

While making the dataset, Attack-Network and Victim-Network, totally were isolated two systems, were de-marked and executed by Sharafaldin H. et al [12]. They gathered information from July 3, 2017, to July 7, 2017, for the dataset.

B. SVM

Statistical learning and curved advancement, in view of the guideline of basic structural risk minimization of Support Vector Machine (SVM) calculations. Vapnik et al created SVM as an answer for various issues. For instance, it tends to be utilized in a wide range of territories, for example, learning, design recognition, relapse, order, and investigation.

SVM is a regulated learning technique since it utilizes labeled data in a dataset as information. The quantity of output classes changes relying upon the dataset. For instance, two classes of data information are created when a dataset of two classes is given as the info. Along these lines, the examples given as the info are sorted by these classes. During training preparation procedure, a model is made by the information dataset and order is performed by utilizing the model.

C. Random Forest

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

Random Forest is a great algorithm to train early in the model development process, to see how it performs and it’s hard to build a “bad Random Forest”, because of its simplicity. This algorithm is also a great choice, if you need to develop a model in a short period of time. On top of that, it provides a pretty good indicator of the importance it assigns to your features.

Random Forests are also very hard to beat in terms of performance. Of course you can probably always find a model that can perform better, like a neural network, but these usually take much more time in the development. And on top of that, they can handle a lot of different feature types, like binary, categorical and numerical.

D. Deep Learning

Deep Learning algorithms allow to extract remove includes automatically from a given dataset and they comprise of a successive layer design. Applying non-linear transformation to the consecutive layer structure comprise the premise of Deep learning algorithm. Expanding the quantity of layers will build the multifaceted nature of nonlinear changes to be built. Deep learning algorithm get familiar with the abstract properties of the information got in the last layer from its dynamic portrayals gained at numerous dimensions. Along these lines, the theoretical properties of the last layer output are obtain by bringing the information into an high-level non linear function.

E. Methodology

The SVM, Random Forest and deep learning algorithm were utilized to distinguish port scan attempts dependent on the CICIDS2017 dataset. The flowchart of the proposed strategy was exhibited in figure 2. Above all else, 692,703 records are taken which consists of 252,672 port scan attempts and 440,031 benign categories that are taken from the dataset and after that these records were standardized. After standardization tests were part into two as 75% preparing information and 25% testing information. What's more, the SVM and profound learning IDS models were made dependent on the preparation information. At long last, the models were tried with test information and the presentation of models was determined nearly. Also, the profound learning IDS model comprise of 5 hidden layers and each layer incorporate the diverse number of neurons, for example, 100, 130, 60, 40 and 5 separately. The input was chosen and utilized as an enactment work in the model. Contingent upon the quantity of neurons and hidden layer model exhibitions were changed In this paper, we chose ideal numbers dependent on the model's exactness. Then again, we didn't matter any component choice calculation for SVM and we utilized all highlights. As a future work, we are going to utilize artificial intelligence to deal with characterize value.

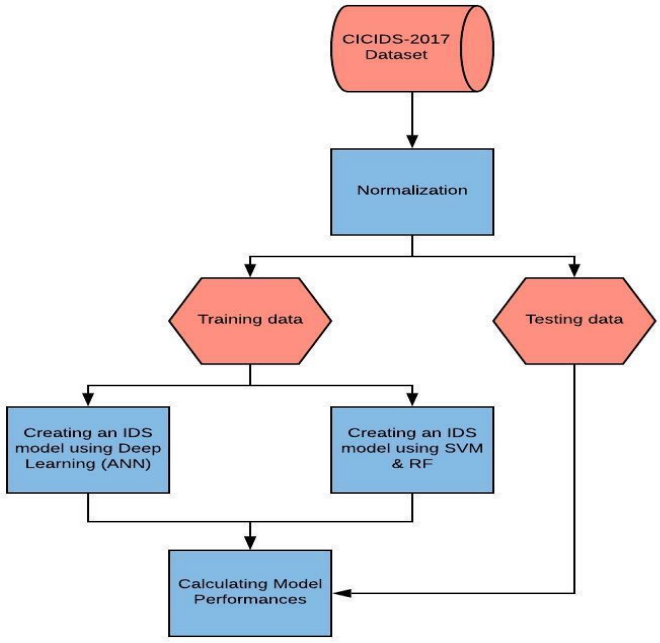


Fig: 2 - Flowchart of the method

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 RF, 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 75% training and 25% 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.

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

TABLE I PERFORMANCE METRICS

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.

IV. EXPERIMENTAL RESULTS

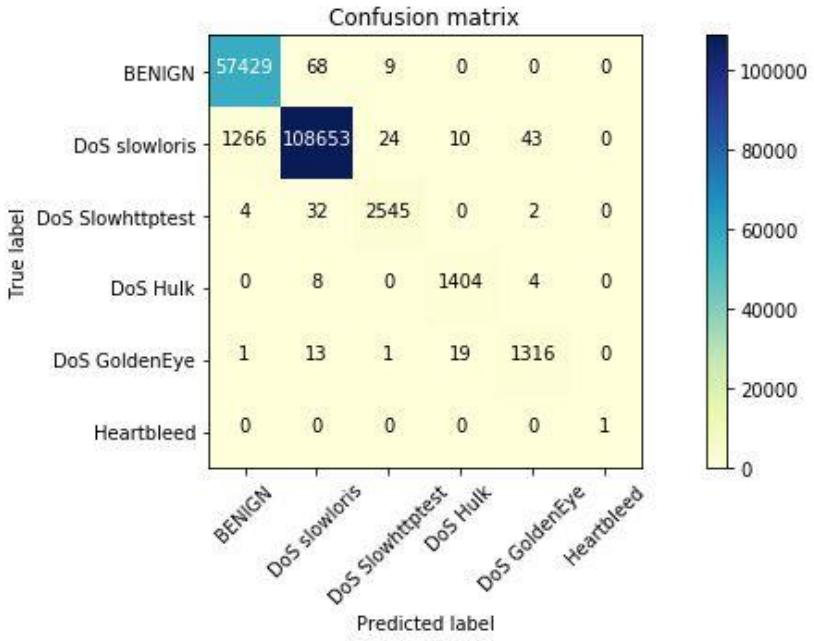
The PC which has Intel(R) Core(TM) i5-6200U CPU @2.30 GHz, 8 GB Ram limit was utilized for analyses. We utilized the CPU anyway we are thinking about to apply GPU as a future work. 692,703 records, which were taken from the standardized dataset, were isolated into two sets with 75% preparing and 25% testing proportions, for example, 519,527 records for preparing and 173,176 records for testing. The profound learning model was prepared in 10 spans and execution estimation of the RF, SVM and profound learning models exhibited in (Table II).

TABLE II
PERFORMANCE METRICS OF USED CLASSFICATION TECHNIQUE BASED ON CICIDS2017 DATASET.

Method	Accuracy	Precision	Recall	F1 score
Deep Learning	**0.9887**	**0.98**	**1.00**	**0.99**
SVM	**0.7219**	**0.79**	**0.72**	**0.66**
Random Forest	**0.9920**	**1.00**	**1.00**	**1.00**

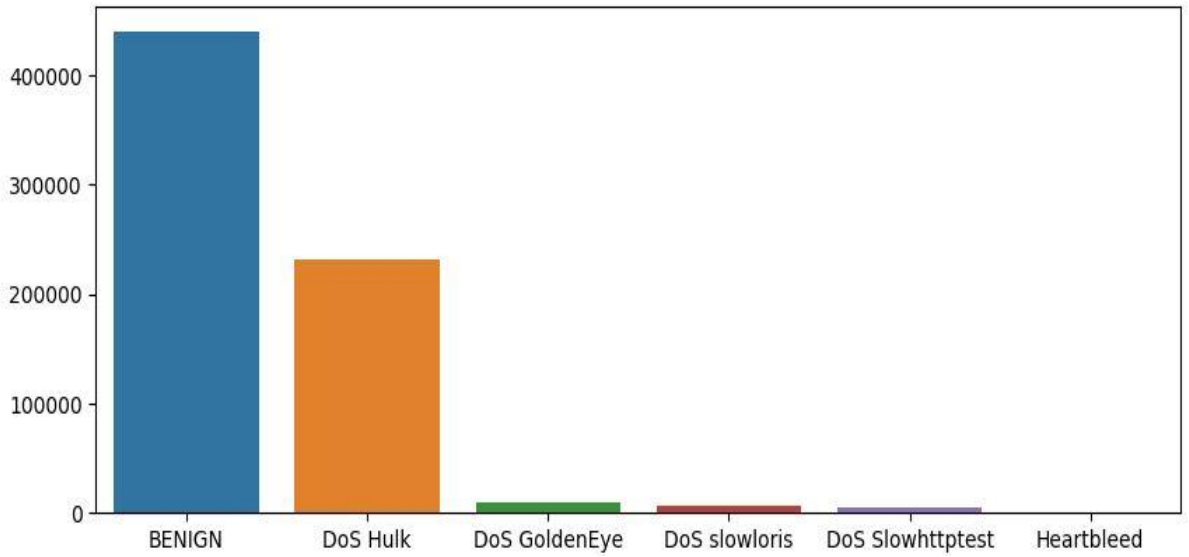
Table IV shows the accuracy, recall, precision and F1 score rates of the IDS models which were developed by using deep learning Random Forest and SVM. Deep learning and RF achieved a higher success than SVM.

The Predicted label over True label (Performance) in Deep Learning model on calculation bases is shown by a below outcome confusion matrix.

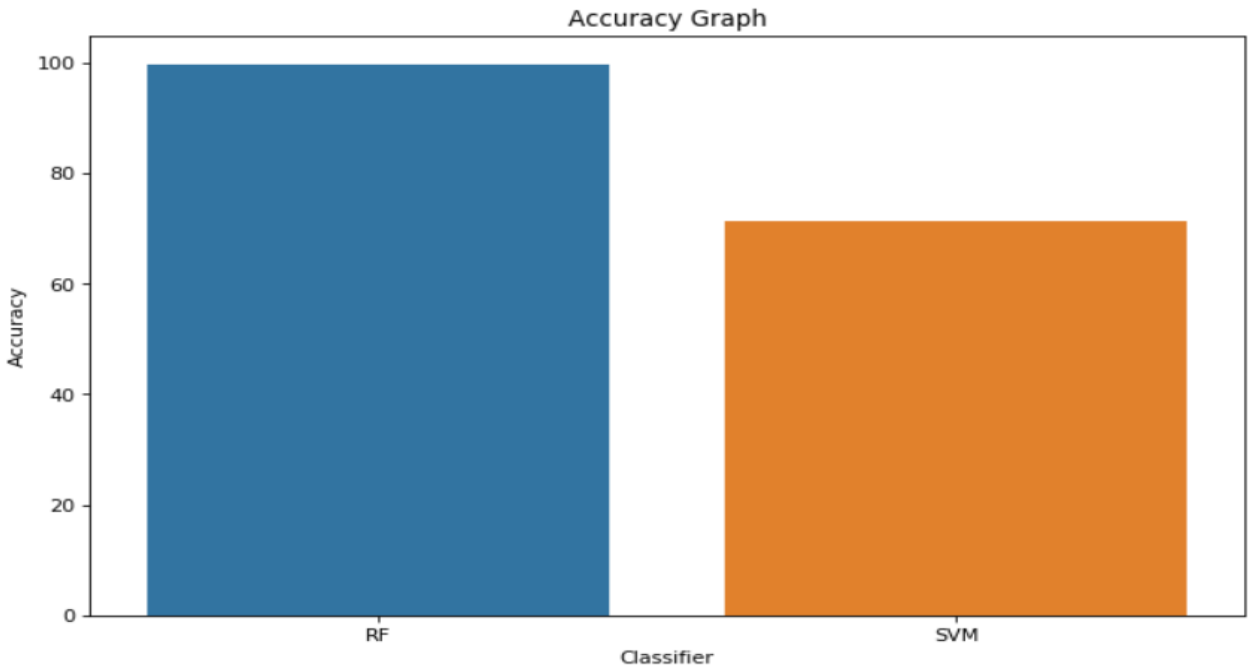


ANN Confusion Matrix

The accuracy performed in this paper by the SVM model is the outcome bar plot graph shown below. The graph consists of port scan attacks like Benign, Dos Hulk, Dos golden Eye, Dos Slowloris, Dos Slowhttptest, heartbleed as labeled in input data set. The performance comparison of RF and SVM on same dataset CICIDS2017 is also defined by the outcome of bar plot graph shown here.



SVM Accuracy Graph



RF & SVM Accuracy Graph

V. CONCLUSION AND FUTURE WORKS

In this paper, execution estimations of Support Vector Machine, Random Forest and Artificial Neural Network, calculations dependent on exceptional CICIDS2017 dataset were introduced similarly. Results demonstrate that the ANN & RF calculation performed altogether preferred outcomes over SVM. We are going to build complete AI model by using Reinforcement Learning to act not only port scan but also other types of attacks later on.

REFERENCES

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, "Port scanning techniques and the defense against them," SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, "Bilgi g ¨uvenli ˘gi sistemlerinde kullanilan arac¸larin incelenmesi," in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

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

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

[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.

[7] N. Moustafa and J. Slay, "The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems," in Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 2015 4th International Workshop on. IEEE, 2015, pp. 25–31.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, "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.

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

[10] M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component analysis for ids," in IEEE International Conference on Communication and Electronics Systems, 2016, pp. 1–5.

[11] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in ICISSP, 2018, pp. 108–116.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, "Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm," in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.

[14] N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, "Distributed abnormal behavior detection approach based on deep belief network and ensemble svm using spark," IEEE Access, 2018.

[15] P. A. A. Resende and A. C. Drummond, "Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling," Security and Privacy, vol. 1, no. 4, p. e36, 2018.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[17] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct," Bone marrow transplantation, vol. 49, no. 3, p. 332, 2014.