Analysis of Machine Learning Algorithms for Intrusion Detection

Gaurav Shyam [1], Jedediah [2], Kaustubh Pandey[3], Mendapara Suraj Vallabhbhai[4], Mrs. ChethanaV[5]

[1], [2], [3], [4] – Final year students, Department of CSE, DSATM

[5] – Assistant Professor, Department of CSE, DSATM

Abstract

With outstanding development in the size of PC organizations and created applications, the measure of the potential harm that can be brought about by dispatching attacks is getting self-evident. Network security technology is crucial in safeguarding governmental and various other industries software infrastructures. Present day intrusion detection applications face complex prerequisites; they should be dependable, extensible, simple to oversee, and have low support cost. Profound neural organizations have exhibited their viability in most AI errands, with intrusion detection included. This work intends to look at the proficiency of AI strategies in intrusion detection frameworks, including artificial neural networks, with the desire for giving a reference to setting up intrusion detection frameworks later on.

Keywords - *Intrusion Detection, Neural Networks, Machine Learning, Infiltration*

I. INTRODUCTION

The internet has become a fundamental part of everyday life; from conveying through texts to messaging and shopping, it has consumed every aspect of life. Moreover every other business and product interact with its customer through the internet and as a result, information and data about the customer is gathered and stored. Protecting this data is a responsibility that the product or business should execute with highest priority. Security measures should be designed and deployed without fail to prevent malware and unauthorized access to the data stored as well as the system itself. However, completely blocking and preventing attacks and breaches is not possible, at least at the present. Intrusion Detection Systems can help minimize the damage caused by these attacks by detecting an impending breach. Its goal is to notify instances of security breaches that may compromise any system. However building a perfect IDS is not an easy task as training these systems will require datasets that are not openly available. Most datasets produced are either confidential or not appropriate. However, the datasets that are available to the public are anonymized and are outdated by current standards. In this paper we look to identify the best machine learning algorithm for intrusion detection.

II. AVAILABLE DATASETS

In this part, we investigate and assess a couple of the openly accessible IDS datasets since 1998 to exhibit their deficiencies and issues that mirror the genuine requirement for a complete and solid dataset. **KDD'99 (University of California, Irvine 1998-99):** This dataset has countless repetitive records and is studded by information defilements that prompted slanted testing results. This dataset has countless repetitive records and is studded by information defilements that prompted slanted testing results. **CAIDA (Center of Applied Internet Data Analysis 2002-2016):** In this organization, there are three different datasets which are available, the CAIDA OC48, which contains different forms of data that were notices on an OC48 link, the CAIDA DDOS, which contains the DDoS attack traffic split of 5 minute pcap files, lastly the CAUDA's Equinix-Chicago monitor on the fast-paced Internet backbone. **Kyoto (Kyoto University 2009):** This dataset has been made through honypots, so there is no cycle for manual naming and anonymization, however it has a restricted perspective on the organization traffic on the grounds that lone assaults coordinated at the honeypots can be noticed. It has ten additional highlights, for example, IDS Detection, Malware product Detection, and Ashula Detection than past accessible datasets which are helpful in NIDS examination and assessment. **ISCX2012 (University of New Brunswick 2012):** In this dataset there are two profiles: the Alpha-profile which did

different multi-stage assault situations, and the Beta-profile, which is the favorable traffic generator and creates reasonable organization traffic with foundation commotion. It incorporates network traffic for FTP, POP3, SMTP, SSH, IMAP, and HTTP conventions with full bundle contents.

III. METHODOLOGY

In this work, different AI algorithms like KNN, RF, MLP, Navies Bayes are tried against the ISCX2012 dataset.

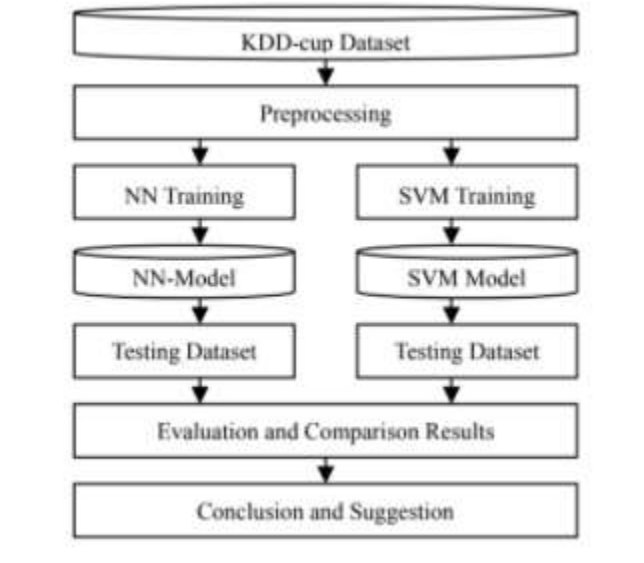

Fig. 1 Architecture of proposed system

1. K-nearest neighbor

k-NN is called example based learning, and it is unique in relation to the inductive learning approach. In this way, it doesn't contain the model preparing stage, however just hunts the instances of info vectors and groups new occurrences. In this way, k-NN "on-line" prepares the models and discovers k-closest neighbor of the new occasion. It figures the rough distances between various focuses on the info vectors and afterward allots the unlabeled highlight the class of its K-closest neighbors. During the time spent making k-NN classifier, k is a significant boundary and diverse k qualities will cause various exhibitions. In the event that k is extensively colossal, the neighbors utilized for expectation will make an enormous grouping time and impact the precision of forecast.

2. Random Forrest

Random Forest is a well-known AI process which has a place with the managed training method. It tends to be utilized for both Classification and Regression issues in ML. It depends on the idea of gathering realizing, which is a cycle of consolidating numerous classifiers to tackle a mind-boggling issue and to improve the presentation of the model. Random Forrest is a type of classifier which comprises of various of decision trees on various subsets of the given dataset and takes the normal to improve the farsighted precision of that dataset. It predicts yield with high precision, in any event, for the enormous dataset it runs productively. It can likewise keep up precision when a huge extent of information is absent.

3. Multilayer Perceptron

There is at least one layer of neurons involved. It comprises three kinds of layers—the hidden layer, output layer, and input layer. Backpropagation learning algorithm is used to train the neurons in MLP. They are truly adaptable and can be utilized for the most part to take in a planning from contributions to yields. This adaptability permits them to be applied to different sorts of information.

4. Navies-Bayes

There are numerous situations where we know the factual conditions or the causal connections between framework factors. Notwithstanding, it very well may be hard to correctly communicate the probabilistic connections among these factors. As

such, the earlier information about the framework is basically that a few factors may impact others.

Once we have collected the dataset, we do the preprocessing of the dataset. Data preprocessing is a cycle of setting up the crude data and making it appropriate for an AI model. It is the first and significant advance while making an AI model. While making an AI project, it isn't generally a case that we stumble upon spotless and perfectly laid out data. And keeping in mind that doing any activity with data, it is obligatory to clean it and place it in an organized manner. So for this, we use data preprocessing tasks. In the event that the missing qualities in a section or highlight are mathematical, the qualities can be credited by the mean of the total instances of the variable. Mean can be supplanted by middle if the component is suspected to have exceptions. For an all-out element, the missing qualities could be supplanted by the method of the segment. The significant disadvantage of this strategy is that it lessens the fluctuation of the credited factors. This technique likewise diminishes the connection between the ascribed factors and different factors on the grounds that the credited qualities are simply gauges and won't be identified with different qualities intrinsically. After preprocessing and filtering of dataset, we run the above mentioned machine learning algorithms and on the filtered dataset. Upon successfully running the algorithms, we do the evaluation of the algorithms.

The evaluation metrics that we will be using average accuracy rate, F- measure, Recall, and Precision.

Precision (Pr): It is the proportion of accurately grouped assaults streams (TP), before all the arranged streams (TP+FP).

Recall (Rc) or Sensitivity: It is the proportion of accurately ordered assault streams (TP), before totally created streams (TP+FN).

F-Measure (F1): It is a harmonic mean of the precision and recall into a single quantity.

Average Accuracy: It is defined as the ratio of the correctly classified data points to the total number of data points.

$$Pr = \frac{TP}{TP+FP}, Rc = \frac{TP}{TP+FN}, F1 = \frac{2}{\frac{1}{Pr}+\frac{1}{Rc}}$$

$$AverageAccuracyRate = \frac{TP+TN}{TP+FN+FP+TN}$$

IV. CONCLUSION

In this paper, we have evaluated the most popular machine learning algorithms in the domain of detecting intrusions and attacks on a system. Using a publicly available dataset ISCX 2012 provided by the Canadian Institute for Cybersecurity, we have extracted the traffic features of a network for 7 days. This data consists of both normal and malicious activity over the network, and we have only evaluated the performance of the algorithms. Having calculated the precision value, f1 score, recall and average accuracy rate for each algorithm we are yet to classify them to specify which algorithm is the most suitable and accurate.

However, the dataset used is not suitable to train, implement and build an IDS. Having a more diverse, updated and reliable dataset is one of the main concerns for every researcher working in this field. Most publicly available datasets lack the diversity and volume of traffic, usually have anonymized packet information, lack metadata and feature set and very few variety of attack data. The future work will focus on obtaining a more reliable dataset either by creating an entirely new one or getting access to one. The possibility of combining two or more algorithms to obtain accurate results is something that could be researched in the coming years.

V. REFERENCES

[1] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin, “Intrusion detection by machine learning: A review” Taiwan, 2009.

[2] Alireza Osareh and Bita Shadgar, ”Intrusion Detection in Computer Networks based on Machine Learning Algorithms” : IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008, Shahid Chamran University, Ahvaz, Iran.

[3] Zheng Wang, *Deep Learning-Based Intrusion Detection With Adversaries*. National Institute of Standards and Technology, Gaithersburg, USA, 2018..

[4] Iman Sharafaldin, Arash Habibi Lashkari , and Ali A. Ghorbani “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization” In proceedings of the 4th International Conference on Information Systems Security and Privacy(ICISSP), Portugal, 2018