Artificial Intelligence based Network Intrusion Detection with Hyper-Parameter Optimization Tuning on the Realistic Cyber Dataset CSE-CIC-IDS2018 using Cloud Computing

V. Kanimozhi and T. Prem Jacob

***Abstract*—One of the latest emerging technologies is artificial intelligence, which makes the machine mimic human behavior. The most important component used to detect cyber attacks or malicious activities is the Intrusion Detection System (IDS). Artificial intelligence plays a vital role in detecting intrusions and widely considered as the better way in adapting and building IDS. In trendy days, artificial intelligence algorithms are rising as a brand new computing technique which will be applied to actual time issues. In modern days, neural network algorithms are emerging as a new artificial intelligence technique that can be applied to real-time problems. The proposed system is to detect a classification of botnet attack which poses a serious threat to financial sectors and banking services. The proposed system is created by applying artificial intelligence on a realistic cyber defense dataset (CSE-CIC-IDS2018), the very latest Intrusion Detection Dataset created in 2018 by Canadian Institute for Cybersecurity (CIC) on AWS (Amazon Web Services). The proposed system of Artificial Neural Networks provides an outstanding performance of Accuracy score is 99.97% and an average area under ROC (Receiver Operator Characteristic) curve is 0.999 and an average False Positive rate is a mere value of 0.001. The proposed system using artificial intelligence of botnet attack detection is powerful, more accurate and precise. The novel proposed system can be implemented in n machines to conventional network traffic analysis, cyber-physical system traffic data and also to the real-time network traffic analysis.**

***Index Terms*—Artificial Intelligence, AWS, CSE-CIC-IDS2018, hyper-parameter optimization and realistic network traffic cyber dataset.**

I. Introduction

THE objective of network intrusion detection is at identifying and monitoring malicious activities. Most of the current IDSs can be partitioned into two fundamental classes. They are intrusion detections based on signature and based on anomaly IDS. An IDS based on signature detects by comparing the already known attacks with the incoming network traffic tries to detect the intrusions, that are stored in the database as signatures. Existing attacks are well detected by IDS, but it often fails to detect novel (unseen) attacks. The next category is called IDS based anomaly. The normal traffic is modelled by the IDS models through learning patterns in the training phase. The deviations from these learned patterns are labelled as anomaly or intrusion. The implementation of real-time IDS based on anomaly is a herculean task because of the rapid increase in the network traffic behavior and very limited availability of computational resources (computation time and memory)[1-5].

There is another challenge, and that is the risk of over fitting due to the high dimensional feature space and the model complexity of IDS. Artificial Intelligence (AI) based techniques play a crucial role in the development of IDS and has more advantages over other techniques. There is no appropriate and well-defined technique to solve the anomaly detection problems[6-8].

The scheme of the proposed system can facilitate the higher understanding of various intrusion detections during which analysis has been made in the sphere of IDS. They are helpful for those that have an interest in applications of AI-based techniques to IDS and connected fields. In this paper, we proposed an experimental approach of Artificial Neural Networks with hyper-parameter optimization on the realistic new IDS cyber dataset (cse-cic-ids2018) included most of the up-to-date attacks (PCAP) along with labeled flows covering more than 80 features (CSV) which obtained through cloud computing on AWS services for intrusion detection in order to provide more accurate accuracy[9-12].

The rest of the paper sectioned as below. Section II contains the background of the work and previous work. Section III and IV describes the methodology and implementation of the work. Section V discuss about the results. At last, Section VI concludes the paper with conclusion.

V. Kanimozhi, Research Scholar, with the Department of Computer Science, Sathyabama Institute of Science and Technology, Chennai, India (e-mail: kanimv@yahoo.co.in)

Dr. T. Prem Jacob, Associate Professor, with the Department of Computer Science, Sathyabama Institute of Science and Technology, Chennai, India (e-mail: premjac@yahoo.com)

978-1-5386-7595-3/19/$31.00 ©2019 IEEE

II. Background And Previous Work

A lot of research work carried out in Intrusion Detection Methods Either Intrusion Detection On Host (HIDS) Or Intrusion Detection on a Network (NIDS) and also on Artificial Intelligence, but there is no comprehensive reliable cyber dataset which covers both contemporary and modern-day attacks for network intrusion detection system[13,14]. According to Alex Shenfield and his co-authors stated that the studies did an offline technique for recognizing shellcode designs inside data [1].

Networks are more vulnerable day by day due to modern attacks. In this proposed system, we make use of realistic latest cyber dataset which compromises of both existing attacks and zero-day attacks by Canadian Cybersecurity which is obtained through cloud computing[15,16].

III. Methodology

A. Botnet

A botnet is an attack which is coined from two words "Robot" and "Network". It is a network which can be operated or commanded by remotely controlled computers. Moreover, it is nothing but a malware that makes the system or server to be controlled and commanded remotely by an operator.

Botnet performs various criminal and malicious activities like stealing information especially in the banking and financial sectors by logging and grabbing customers information. It hijacks the confidential information like login names, user passwords, and other credential information. Some of the botnet attacks techniques described as follows.

Crypto-Locker ransomware - It is a luxurious software which attacks the window operating system by encrypting all the files in the user system with RSA-2048 public key. It claims hefty ransom in order to decrypt the file. It is astounding fact that the virus earned $30 million in hundred days.

Cyber apocalypse - It is another malware and poses a serious threat to networks which produces the impact of an army of bots. Eg of bots. Zeus, Ares etc. Denial-of-service attacks are launched by Botnets of zombie computers which can be propagated through Drive-by-downloads and spam emails.

Credential stuffing is a malicious activity which handles the automated injection attack by making usage of botnets in it to access the online services by stealing the significant credentials. Researchers from Akamai reported the fact that 30 hundred million malevolent login endeavors were created between Gregorian calendar month 2017 to June 2018 from the states of the U.S., Russia, and Vietnam.

B. Artificial Neural Networks

Artificial Neural Network models are a structure gaining knowledge from the machine that endeavors to mirror the learning example of natural neural systems ie.biological system. Natural neural systems work in the sense that the dendrites receive inputs which are said to be presented in the interconnected neurons of the human brain. Based on these inputs, through an axon to another neuron, they produce an output signal. We will attempt to imitate this procedure utilizing Artificial Neural Networks (ANN), simply refer to as neural systems starting now and into the foreseeable future. Neural systems are the establishment of profound learning. It is a subset of machine learning in charge of the absolute technical advances today!

C. Multi-layer Perceptron tuning with hyper-parameter optimization Classifier Model

The realistic cyber dataset is preprocessed and all the object features are converted into numeric features and we trained the model for Artificial Intelligence in the training phase and finally we tested to find the detection accuracy in the testing phase. We used Multi-Layer Perceptron (MLP) to build the proposed Artificial Intelligence. Perceptron is delicate to attribute or feature scaling. Therefore, scaling your data should be advisable.

A Perceptron has the subsequent: one or extra inputs, a bias, a function of activation, and a resultant output. Inputs (ie.,80 features of this IDS2018 dataset) are received by the perceptron, applies a few weight, and the output (attack or normal) is produced by way of the activation unit which receives the weighted inputs. The neural network can be modeled by adding perceptrons layers together to form Multi-layer perceptrons of our proposed framework.

As for hyper-parameter optimization, GridSearchCV Optimization technique is used. Tuning a neural network for optimization is a herculean task and it is a lengthy process. The hyper-parameters which are considered for tuning are alpha that could be a comparison of various values of regularization parameters and another parameter for tuning is hidden layer sizes. It operates on parallel and can be iterated, with 10-fold cross-validation. We model our neural network by starting with two layers [1].

Solver has been picked in this model as 'lbfgs'. And try to find alpha parameter using L2 regularization. Better prediction and accuracy will not be generated without regularization method. In our proposed framework, we tend to distinguish the classification either “Benign” or “Malicious” supported the output.

High-quality F1 rating : 0.9991678456370812
Fine parameter : 'alpha': 1e-05,'hidden_layer_sizes': (9, 4)

IV. Implementation

A. CSE-CIC-IDS2018

We built an MLP Classifier model on realistic cyber defense dataset by Canadian Institute for Cybersecurity (CIC) on AWS (Amazon Web Services). *Datasets* by *CIC* and ISCX are used around the world for security testing and malware prevention. Knowledge on AWS is a must for accessing that dataset which is stored in Resource type **-S3 Bucket** and Amazon Resource Name(ARN) is *arn:aws:s3:::cse-cic-ids2018 and also AWS Region Ca-central-1under License*[17].

It consists of a detailed description of intrusions at the side of abstract distribution patterns for the subsequent: applications, network protocols, and other sublevel network entities. The very last dataset includes seven distinct attack eventualities: brute-force, heartbleed, botnet, dos, ddos, web assaults, and infiltration of the network from inside. The attacking infrastructure consists of 50 machines. The victim agency has five departments and consists of 420 machines and 30 servers. The dataset consists of the captures community site visitors and gadget logs of every machine, along with eighty attributes extracted from the captured traffic the usage of CICFlowmeter-v3[2].

B. Creating an Artificial Neural Network with Anaconda, Jupyter Notebook and SciKit- Learn

To build this Artificial Neural Network, we use Anaconda 3.0 and the latest Scikit version 0.19.1 and Pandas version 0.23.1 in Jupyter Notebook. It can be installed through pip or Miniconda (Package Manager).

C. Receiver Operating Characteristics Curve

Receiver Operating Characteristics curve is utilized to picture the execution of multi-dimensional data classification. It is being considered as one of the most prominent evaluation metrics for evaluating any classification model's accuracy. It is also referred to as AUROC(Area Under the Receiver Operating Characteristics)

Let's start the proposed artificial Intelligence model. To get the whole evaluation metrics, I have created two functions. The calculate_auc function also produces ROC. To make an outline for the basic layout of execution measurements, and that has been executed by pandas.

V. Results

A. ROC Curve

The curve generated in Fig. 1 when True positive versus against False Negative rate at various threshold points and the curve implies how well the binary classifier discriminated between two different classes i.e., Benign or malicious. Theclassifier model runs a sample of 1048575 records with 80 features and optimize it with 10 Fold Cross Validation to produce the ROC curve in Fig. 1.

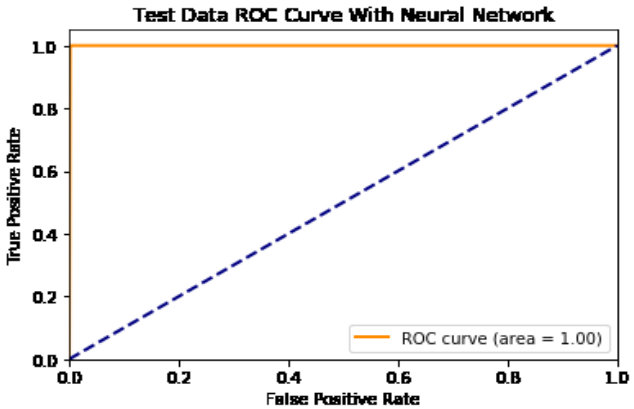

Fig. 1. ROC CURVE

B. AUC SCORE

It is the region under the roc curve, and it outlines the overall executed performance of the binary classifier. Higher the score, better the classifier model performance.

AUC SCORE : 0.9991680

C. Confusion Matrix

It gives insights of the number of positive and negative predictions and also summarizes the count of normal and malicious attacks in this model and the below graph is shown with samples how 100% it identifies the normal and malicious botnet attacks. So overall confusion Matrix outperforms the evaluation metrics of this model which is shown in Fig. 2.

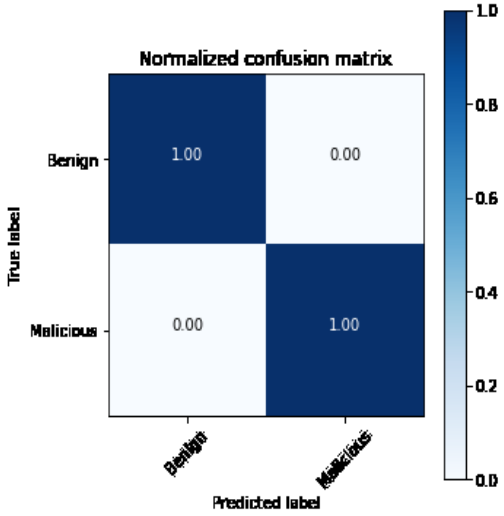

Fig. 2. Confusion Matrix of Neural Network

D. Classification Report of Neural Network Model

The Classification Report of our proposed Artificial Intelligence is as shown below in Table I and also the Accuracy score is given below.

TABLE I
CLASSIFICATION REPORT OF NEURAL NETWORK

Training Data Performance Metrics				
Accuracy	Precision	Recall	F1	AUC
1.0	1.0	1.0	1.0	1.0
Test Data Performance Metrics				
Accuracy	Precision	Recall	F1	AUC
0.9997	1.0	1.0	1.0	1.0

Artificial Intelligence Model Training Accuracy: 1.0
Artificial Intelligence Model Testing Accuracy: 0.99975

E. Default MLP Classifier Model Comparison

If the version has no longer been set by way of any parameter, the default value of alpha is 0.0001 and hundred neurons is the size of a single layer. You may envision the score of accuracy and the power of optimization can be realized.

Artificial Intelligence Model Training Accuracy: 0.99983
Artificial Intelligence Model Testing Accuracy: 0.9995

VI. Conclusion

The proposed system can be extended to detect all other remaining classes of attacks in this realistic dataset which includes all real-time and existing attacks. The framework used in this artificial Intelligence Scikit learn framework optimization is based on Central Processing Unit, not on Graphics Processing Unit, the optimization may be powerfully tuned by other such frameworks like Google's open sourced Tensor Flow. The performance issue is a common task when we come across pandas to work with larger data (100 gigabytes to multiple terabytes), but Spark is an open-sourced Apache Framework used for big data processing can handle parallel computing with massive datasets, ranging from 100 gigabytes to multiple terabytes across clustered computers.

References

[1] Alex Shenfield, David Day, and Aladdin Ayesh, "Intelligent intrusion detection system using artificial neural networks," vol. 4, no.2, pp. 95-99, June 2018.

[2] Iman Sharafaldin, ArashHabibiLashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.

[3] D. Stiawan, A.H. Abdullah, and M.Y. Idris, "The trends of intrusion prevention system network, in 2010" 2nd International Conference on Education Technology and Computer, vol. 4, pp. 217-221, June 2010.

[4] Singh R., Kumar H., Singla R.K., and Ketti R.R. "Internet attacks and intrusion detection system: A review of the literature"Online Inform. Rev., 41 (2), pp. 171-184, 2017.CrossRefView Record in ScopusGoogle Scholar.

[5] Liao H.-J., Lin C.-H.R., Lin Y.-C., and Tung K.-Y. "Intrusion detection system: A comprehensive review" Network Computing. Appl., Rev., 36 (1), pp. 16-24,2013. [Online]. Available https://www.kdnuggets.com/2016/10/beginners-guide-neural-networks-python-scikit-learn.html. [Acessed:14-SEP-2018]

[6] Zhang G.P. "Neural networks for classification: A survey" IEEE Trans. Syst. Man Cybern. C, Rev., 30 (4), pp. 451-462, 2000.

[7] Wu J., Peng D., Li Z., Zhao L., and Ling H. "Network intrusion detection based on a general regression neural network optimized by an improved artificial immune algorithm."Rev.,10 (3), 2015. [Online] Available https://www.ncbi.nlm.nih.gov/pubmed/25807466 [Acessed:14-SEP-2018].

[8] Rosenblatt F. "The perceptron: A probabilistic model for informationstorage and organization in the brain" Psychol. Rev., 65 (6), pp. 386-408, 1958.

[9] Gulshan Kumar. "The use of artificial intelligence based techniques for intrusion detection: a review", Artificial Intelligence Review, 09/04/2010.

[10] Antonia Nisioti, Alexios Mylonas, Paul D. Yoo, Vasilios Katos. "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods", IEEE Communications Surveys & Tutorials, 2018.

[11] Monowar H. Bhuyan, Dhruba K. Bhattacharyya, Jugal K. Kalita. "Network Traffic Anomaly Detection and Prevention", Springer Nature, 2017.

[12] JChristina Ting, Richard Field, Andrew Fisher, Travis Bauer."Compression Analytics for Classification and Anomaly Detection within Network Communication", IEEE Transactions on Information Forensicsa nd Security, 2018.

[13] Sang-Jun Han, Sung-Bae Cho. "Rule-based integration of multiple measure-models for effective intrusion detection", SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat.No.03CH37483), 2003.

[14] Ghorbani A., Lu W., and Tavallaee M., 2010, "Network Intrusion Detection and Prevention: Concepts and Techniques", Springer Science, LLC.

[15] Pervez M. S. and Farid D. M., "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014), Dhaka, 2014, pp.1-6.

[16] Engen, Vegard. Machine learning for network based intrusion detection: an investigation into discrepancies in findings with the KDDCUP'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data. Diss. Bournemouth University,2010.

[17] License: http://www.unb.ca/cic/datasets/ids-2018.html [Acessed:14-SEP-2018]