Applying Machine Learning on Intrusion Detection System Dataset

Fekadu Yihunie
School of Computer Science and Engineering
Sacred Heart University
Fairfield, CT
yihunief@sacredheart.edu

Eman Abdelfattah, Amish Regmi
School of Theoretical & Applied Science
Ramapo College of New Jersey
Mahwah, NJ
eabdelfa@ramapo.edu, aregmi@ramapo.edu

***Abstract*—The tremendous growth of internet-based traffic exposes corporate networks for wide variety of vulnerabilities. Intrusive traffics are affecting the smooth operation of network infrastructure by consuming corporate resources and time. Efficient way of protection, identification and mitigation from intrusive incidents enhance productivity. Intrusion Detection system (IDS) is one of the key components of network traffic security. IDS solution can be host based or network based to fully oversee intrusive traffic in the network. Efficient automated detection techniques of anomaly traffic are improving over time. This research aims to find the best classifier that detects anomaly traffic from NSL-KDD dataset with high accuracy level and minimal error rate by experimenting with different machine learning methods. Five binary classifiers: Stochastic Gradient Decent, Random Forests, Logistic Regression, Support Vector Machine and Sequential Model in Keras are tested and validated to come up with the result. The results demonstrated that Random Forest Classifier outperformed the other four classifiers with and without passing through data normalization process.**

Index Terms – IDS, NSL-KDD;

I. INTRODUCTION

Nowadays internet-based applications and dependency of cloud-based services are increasing exponentially. Organizations are focusing on their core businesses and moving their IT services in the cloud. Many other reasons push companies to rely on internet-based services. Similarly, the growth of malicious traffic is rapidly growing. Targeted companies were attacked in different techniques by organized cyber terrorist and script kiddies. Protecting, detecting and managing intrusive incidents are challenging and costly, as organizations strive to comply with different standards.

Well secure network infrastructures are recommended to have firewalls, intrusion detection and prevention systems, and web content and URL filtering devices to protect internal systems from attacks lunched by intruders. The advancement of attacking techniques and the intelligence of organized criminals make it difficult to fully protect sensitive information from theft, disclosure and denial of service attacks. Researchers are studying various machine learning methods to improve the efficiency of intrusion detection systems.

This paper targeted intrusion detection system analysis with various machine learning binary classifiers by using NSL-KDD dataset. Although NSL-KDD dataset is not a perfect representative of existing network traffics, but it is used in this research because of the lack of public datasets, [3].

This paper is organized as follows: section II presents intrusion detection related work, section III describes the contents of NSL-KDD dataset, section IV presents experimental results and analysis of various classification techniques. Finally, section V offers the conclusion and future work.

II. RELATED WORK

Laheeb *et al.* studied a comparison for intrusion detection dataset KDD99 and NSL-KDD based on Self Organization Map (SOM) artificial neural network [1]. They used unsupervised artificial neural network in hierarchical anomaly intrusion detection system. SOM neural nets employed for detection and separation of normal traffic from the attack traffic. The paper has also evaluated the efficiency of SOM in anomaly intrusion detection.

Shilpa *et al.* researched on feature reduction using principal component analysis for effective anomaly–based Intrusion Detection on NSL-KDD dataset [2]. They reduced the number of features in NSL-KDD dataset that are irrelevant and redundant for anomaly detection process. They applied hybrid principal component analysis neural network algorithm to effectively detect attacks by reducing computer resource utilization.

S. Revathi *et al.* analyzed NSL-KDD dataset using various machine learning techniques for intrusion detection system. The analysis focused on selected NSL-KDD datasets to get a good analysis on various machine learning techniques [3]. The Random Forest classification algorithm had a highest accuracy rate as per their experimental result compared to other classification algorithms.

L.Dhanabal *et al.* studied on NSL-KDD dataset for intrusion detection system based on classification algorithms [4]. The paper analyzed and used NSL-KDD dataset to study the effectiveness of various classification algorithms in detecting anomaly network traffic patterns. They analyzed the relationship of the protocols in the network protocol stack with the attacks to generate anomalous network traffic.

Hee-su *et al.* examined feature selection for intrusion detection using NSL-KDD dataset [5]. They identified

important selected input features in building IDS with computationally effective and efficient manner. In the paper, the performance of standard feature selection methods evaluated, and the authors proposed a new feature selection method.

Rowayda *et al.* investigated effective anomaly Intrusion Detection System based on a new hybrid algorithm named neural network with Indicator Variable and Rough Set for attribute Reduction (NNIV-RS) [6]. The experimental results showed the proposed NNIV-RS algorithm has better and robust representation of data and able to reduce unnecessary features to improve the reliability and efficiency of IDS.

Bhupendra *et al.* analyzed the performance of NSL-KDD dataset using artificial neural network (ANN). The result obtained for both binary class as well as five class classification on attacking types analyzed based on various performance measures. The accuracy of ANN was presented [7].

Ray investigated the effects of architecture on the performance of intrusion detection systems (IDSs) [8]. An equation was formed to find the optimal number of hidden neurons in a multi-layer feed forward neural network (MLFFNN) IDS. This equation can be used to determine the number of hidden neurons to eliminate the lengthy trial and error calculations in case of MLFFNN.

III. DATASET DESCRIPTION

The selection of the dataset highly affects the performance of the algorithm we are appling. The NSL-KDD dataset suggested to solve the inherent problems of the KDDCUP'99 dataset [5]. NSL-KDD dataset has removed redundant records in the train dataset and test dataset to enable classifiers to produce an unbiased result [3].

This paper uses the training dataset and test dataset that are made up of two target values, normal and anomaly. The known attack types are grouped as anomaly traffics while the remaining traffics were categorized as normal traffic. The original NSL-KDD dataset has 41 features and a label. NSL-Preprocessing step was conducted as KDD dataset has three features object values that should be changed to numbers before applying classifiers. The three features are as follows: 'protocol_type' has 3 unique categories, 'service' has 70 unique categories and 'flag' has 11 unique categories.

After one-hot encoding was applied on the dataset, the number of features reached 122 and a label, which is assigned for each instance. The total instances in the dataset are 125,973 that were split into train dataset and test dateset. The train dataset has 100,778 instances and the test dateset has 25,195 instances. Figure 1 and Figure 2 depict the number of normal and anomaly instances count in the train and test datasets.

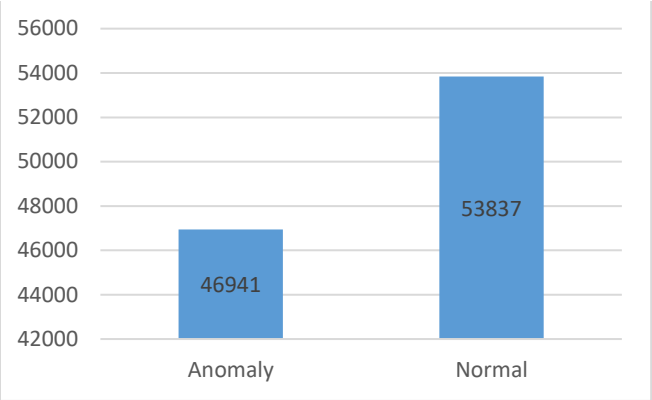

Figure 1. Train dataset target counts

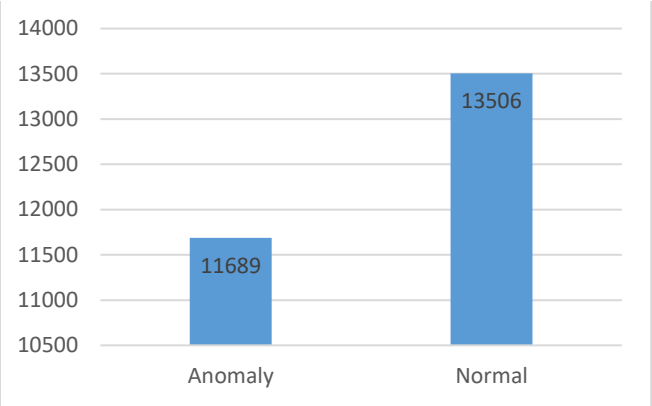

Figure 2. Test dataset target counts

On both datasets the number of anomaly records is lower than the normal. The consistency and fair distribution of instances in the training and test datasets are demonstrated as shown in Figure 1 and Figure 2.

IV. EXPERIMENTAL RESULT AND ANALYSIS

This paper applied different techniques of classification and analyzed the NSL-KDD dataset in numerous ways. Different performance measures were calculated and compared; Precision, Recall, F_1 score, Receiver operating characteristic (ROC) curve. The precision is calculated by dividing the number of true positive (TP) instances over the sum of the number of true positive and false positive (FP) instances. The recall is calculated by dividing the number of true positive instances over the sum of the number of true positive and false negative (FN) instances. The equation for calculating F_1 score is as follows [10]:

F_1 score= 2TP/(2TP+FN+FP)

The classifier can get a high F_1 score only if both recall and precision are high. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR).

Stochastic Gradient Descent (SGD) also known as incremental gradient decent classifier has advantages of handling very large dataset and dealing with training instances independently. This classifier demonstrated poor performance initially because features have the large gaps

between minimum and maximum values. However, by applying standard feature scaling, the problem was solved. The ROC curve for SGD technique is shown in Figure 3.

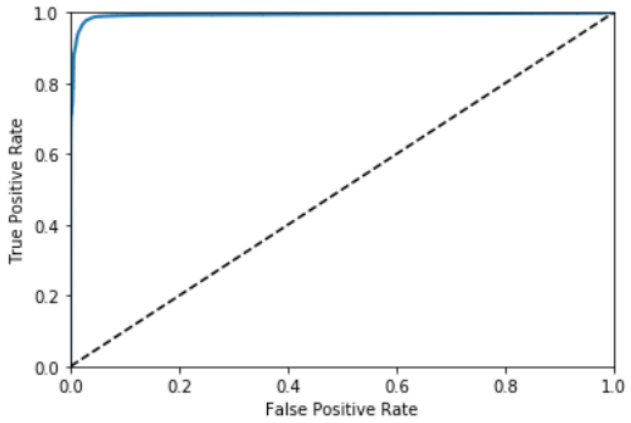

Figure 3. SGD ROC curve

Random Forests classifier works by training many Decision Trees on random subsets of the features, then averaging out their predictions [10]. Random Forests classifier demonstrate good performance. The accuracy level achieved in Random Forests is near to perfection. Figure 4 depicts ROC curve of Random Forests classifier.

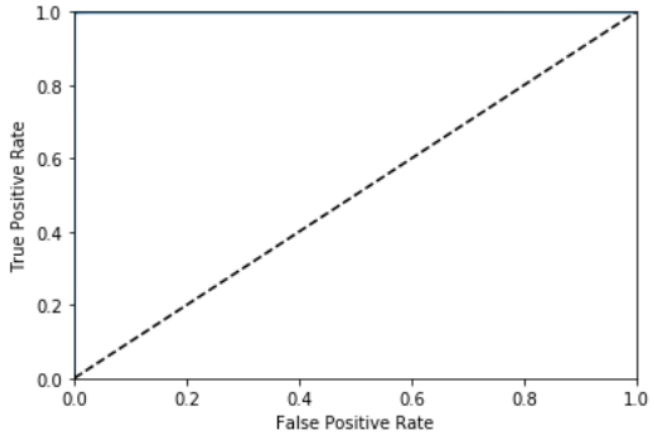

Figure 4. Random Forest ROC curve.

Logistic Regression is one of the regression algorithms that can also be used for classification. Logistic Regression also called Logit Regression is used to estimate the probability that the instance belongs to a particular class [10]. This classifier had less performace compared to the other classifiers applied in the dataset. Figure 5 shows the ROC curve of Logistic Classifier.

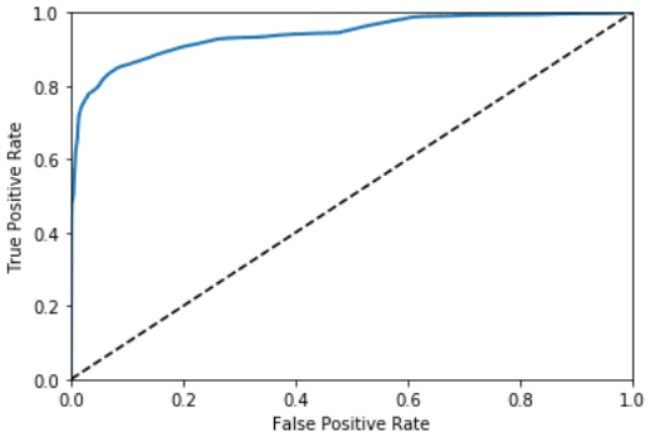

Figure 5. Logistic Regression ROC curve.

A Support Vector Machine (SVM) is a powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification and regression. SVM is well suited for classification of complex but small or medium sized datasets [10]. The result of applying SVM classifier in NSL-KDD dataset demonstrated a good performance and is comparable to the result obtained in case of Random Forests classifier. Figure 6 shows the ROC curve of Support Vector Machine.

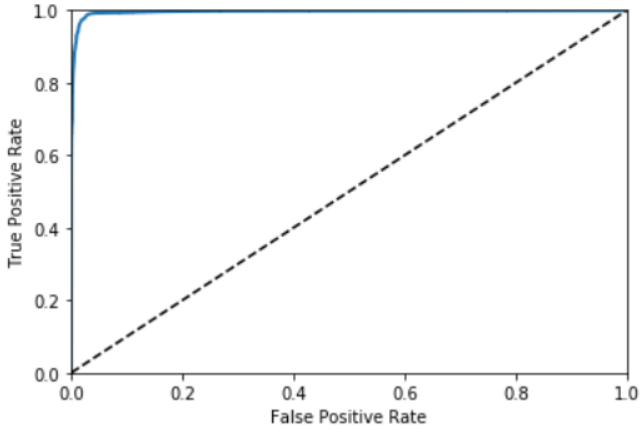

Figure 6. SVM ROC curve.

Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. The recommendation for using Keras is for its easy and fast prototyping of deep learning libraries, through user friendliness, modularity and extensibility. It supports both convolutional neural networks and recurrent networks, as well as combinations of the two, and runs seamlessly on CPU and GPU. The core data structure of Keras is a model, and the simplest type of model is the Sequential model, a linear stack of layers. The input for the model is specified. Before training, the learning method needs to be configured, which is done via the compile method. However, in the experiment conducted Random Forest and SVM outperformed the sequential model in Keras as shown in Figure 7.

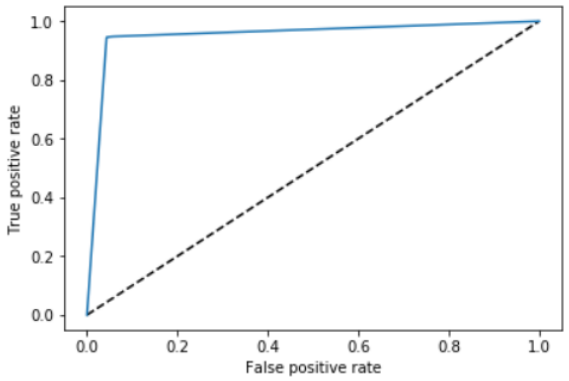

Figure 7. The ROC curve of Sequential Model in Keras

Table 1 includes a summary of precision, recall and F_1 score. Accuracy results for the five classifiers are shown in Figure 7.

Score Type	SGD	Logistic	Random	SVM	Sequential Model
Precision	0.9696	0.8967	0.9992	0.9779	0.9881
Recall	0.9742	0.8507	0.9969	0.9730	0.924
F_1	0.9719	0.8731	0.9980	0.9755	0.95497

Table 1. Score Summary

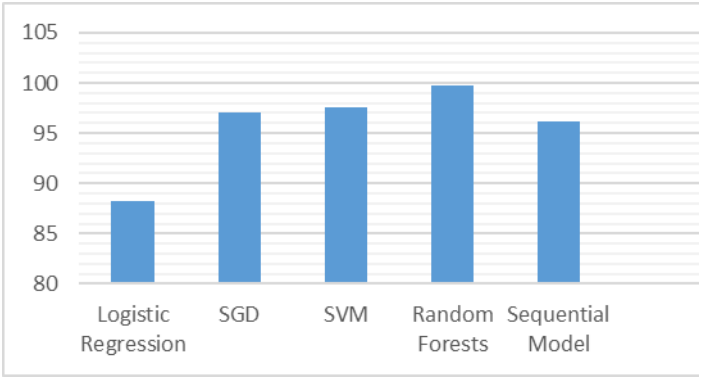

Figure 8. Accuracy results of the five classifiers

V. CONCLUSION AND FUTURE WORK

The comparison of different machine learning models on intrusion detection systems NSL-KDD dataset was conducted. The research has been carried out with five different classification algorithms with and without one-hot encoding. It is obvious that Random Forests algorithm outperformed the other four classifiers. The overall results of Random Forests classifier are near to perfection and outstanding result from earlier published papers was obtained. In our future work, we plan to integrate and analyze various artificial neural networks to classify different class types or attacking techniques in intrusion detection systems dataset.

VI. REFERENCES

[1] Laheeb M. Ibrahim, Dujan T. Basheer and Mahmod S. Mahmod, "A Comparison Study for Intrusion Database (KDD99, NSL-KDD) Based on Self Organization Map (SOM) Artificial Neural Network," Journal of Engineering Science and Technology, Vol. 8, No. 1, pp. 107 – 119, 2013
https://core.ac.uk/download/pdf/25739889.pdf

[2] Shilpa Lakhina, Sini Joseph and Bhupendra Verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD," *International Journal of Engineering Science and Technology*, Vol. 2(6), pp. 1790-1799, 2010
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=09445114AAB380FC3577CEFE20A6BD94?doi=10.1.1.168.1957&rep=rep1&type=pdf

[3] S. Revathi and Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection," International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, ISSN: 2278-0181, December – 2013
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.6760&rep=rep1&type=pdf

[4] L.Dhanabal and Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015
https://pdfs.semanticscholar.org/1b34/80021c4ab0f632efa99e01a9b073903c5554.pdf

[5] Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi and Twae-kyung Park, "Feature Selection for Intrusion Detection using NSL-KDD," Recent Advances in Computer Science, ISBN: 978-960-474-354-4
http://www.wseas.us/e-library/conferences/2013/Nanjing/ACCIS/ACCIS-30.pdf

[6] Rowayda A. Sadek, M. Sami Soliman and Hagar S. Elsayed, "Effective Anomaly Intrusion Detection System based on Neural Network with Indicator Variable and Rough set Reduction," IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
https://pdfs.semanticscholar.org/5293/08f11209427939939dfe2b146fdc151cabb66.pdf

[7] Bhupendra Ingre and Anamika Yadav, "Performance Analysis of NSL-KDD dataset using ANN," Signal Processing and Communication Engineering Systems (SPACES), 2015 International Conference, 2015, Page(s):92- 96

[8] L. Ray, "Determining the Number of Hidden Neurons in a Multi-Layer Feed Forward Neural Network," *Journal of Information Security Research*, vol. 4, no. 2, pp. 63-70, 2013.

[9] Dataset source: https://github.com/jmnwong/NSL-KDD-Dataset

[10] Book: "Hands-On Machine Learning with Scikit-Learn and TensorFlow"