Grenze International Journal of Engineering and Technology, Jan 2020

Anomaly Detection of Computer Networks using Deep Learning Techniques

Vijaya Kumar B P[1] and Chandrakala[2]
[1]HOD and Prof. Dept. of Information Science, Ramaiah institute of Technology Bangalore, India
Email Id: hod_is@msrit.edu
[2]Department of Information Science, Ramaiah institute of Technology Bangalore, India
Email Id: csbadaga@gmail.com

***Abstract*—In daily realistic activities many users are using the Internet. A tremendous increase in the various number of computer network threats compromises the network system that motivate the network anomaly detection system to be relevant and necessary to be implemented using various deep learning approaches for computer network issue. In the network related systems, the detection of such anomalous situations is still challenging. To tackle the issues related to the anomaly situations, deep neural networks with multiple hidden layers is proposed. Anomaly detection based on the user behavior is most essential to secure the machines from unauthorized activities. Anomalies of real time network data packets, net flow records are collected and applied an appropriated deep leaning models to enhance accuracy and speed by using python programming language. Main objective of the project is to identify anomalies in a computer network traffic using deep learning techniques for appropriate dataset. From the experimental results Random forest algorithm overtops support vector machine, Kohonen-self organizing maps and kmeans algorithms on UNSW-ND15 dataset due to the fact that this dataset doesn't have any redundant network connections and connections being fairly distributed across all the classes.**

***Index Terms*— machine learning, deep learning, UNSW-NB15 dataset, Kohonen SOM, Random forest, ANN, SVM.**

I. INTRODUCTION

Abnormal behaviour of the activities can be identified using a technique called as anomaly detection. Anomalies cause's datasets instances are going to perform dissimilarly with respect to other instances present in a dataset [1]. Those instances resulted in finding anomalies. Now a days anomalies can be seen in various sectors like Business, Education, Military and healthcare and many more. To find out all the anomalies machine learning introduced a sub branch called as deep learning which can be used to achieve good flexibility and performance by learning methods represents the data as a nested hierarchy of concepts within artificial neural network. Major goal of anomaly detection technique is to finding out all kinds of instances in a data driven fashion [2]. According to Hawkins Outlier [3] can be defined as it is an observation that diverges significantly from the other observation as to induce wary that it was promoted by different mechanism. With the increasingly blending of society and internet, day by day internet is changing the way of people live, work and study, but numerous security threats people face are becoming more and more serious. In this paper our focus is on the different

Grenze ID: 01.GIJET.6.1.1_2
© Grenze Scientific Society, 2020

anomaly detection techniques. Self-organizing maps are type of artificial neural networks trained using unsupervised learning method to produce low dimensional, discretized representation of input space of training samples called as map which can be used for dimensionality reduction.

A. Scope of the Project

The scope of the project aims effectively finding of the datasets which makes the model much more sufficient. Dataset may include network flow records, data packets or http logs etc. In order to aid this challenge Deep learning is most suitable method of finding out anomalies present in the dataset. This project eventually helps in finding out the anomalies by effective training of the model with various inaccurate conditions. So, this way originally, we have chosen appropriate dataset and performed pre-processing, once it is done, we train the model using suitable method after training the model we apply the deep learning techniques to find out the anomalies in the data set. The model effectively detects the anomalies present in the dataset and shows to the observer. Mainly efficiency and time complexity are the major concern of this project.

B. Problem Statement

To Design, Develop and Implement the training model by using different inputs data. So machine will able to learn the features and extract the anomalies from the data by using deep leaning techniques.

C. Objectives of the project

Main objective of the project is to identify the anomalies in a computer network traffic using appropriate dataset. To design and develop and implement the Anomaly detection model. Dataset must be trained using appropriate deep learning technique. Evaluating of metrics like Precision, recall and Accuracy.

D. Challenges

Challenges are the major basis which immenent the negative impacts on current project.Some of the challengs faced during anomaly detection are :

- choosing appropriate dataset, after choosing dataset tuning of the paremeters which makes project more efficient to get the desired results.
- Model must be trained by taking consideration of less computational efficiency and power.
- Increase of error rate due to dynamically changing the environment.
- Streams of data induces unbounding of data points in anomaly detection.
- Incable of aggrigation of features.

Following sections are literature survey is explained in section 2 and section 3 explaines about methodologies used for implementing the project and final section deals with results and conclsion of the project.

II. LITERATURE REVIEW

As a part of literature survey study of research paper is most relevant to anomaly detection in deep learning. Author in paper[1] proposed deep leaning approach for anomaly detection using recurrent neural network,according his results RNN-IDS model is the strongest model for anomaly detection which resulted in high accuracy in multiclass and binary classification.this task is carried under multiclass classfication on the NSL-KDD dataset. In paper[2] Author presented machine learning approach for automatic detection of anomalies in a person's blood glucose levels, by considering historical observations as a benchmark.In[3] author undertaken to multiclassification of UNSW-NB15 dataset used for computer anomaly detection is implemented using online AODE algorithm. This algorithms helped in updating UNSW-NB15 features it allowed to build a model which adapt the recent computer network anomalies passing over the time. The main objective[4] of this work gives an idea to researchers that what all of the datasets are there for computer network intrusion detectiona and what each of the dataset comprised of like in terms of attacks,features etc. This paper gives detatiled description of dataset instances,classes and attributes also the nature of attributes. In this paper [5] one type of machine learning approach is used to detect the anomalies in the intrusions. So here the creator proposed the peculiarities for three strategies, for example, the Host based IDS, Network Based IDS and alongside that Hybrid IDS. At first here the model must be prepared for various interruptions and by ceaselessly train the model and in the event that we demonstrate another dataset as an information, at that point it tends to be capable discover the abnormalities in the interruption discovery framework. In Thailand the assembling organization produce the generation in the ceaseless stream. While in this production if any error occurred, then the service engineer is only solution to again bring back the erroneous condition to normal condition. Author in paper [6] has profound the machine learning model to correct the error in log files and it gives the report to check the results. The results

are very accurate and diagnosis can be performed by the engineers to evaluate the quickly. Author in [7] presented by considering SVM as major component they proposed their own SVM classification approach and pre-processing of data for better SVM inputs. They designed an enhanced support vector machine for incorporating both supervised and unsupervised scheme which operates without use of labels. Also they have used m-fold cross validation methods for verifying the results and prove that effectiveness of their framework over existing network intrusion detection system. In this paper [9], they have examined the state of art in the anomaly based computer network intrusion detection. Here author has discussed two well know criteria for classifying and evaluating network intrusion detection they are detection strategy and evaluation dataset. Brief description of different datasets are explained.

III. METHODOLOGY

Deep neural network mainly consists of three learning methods, namely supervised learning, reinforcement leaning and unsupervised leaning methods of training a model. Supervised learning is a learning method which maps known input resulted into output which maps from input to output. But in case of unsupervised leaning we would not be knowing targeted output in this leaning we should train the model in order to get desired output.

A. Data set description

Generally, researchers used KDD CUP 1999 dataset for anomaly detection of computer networks. Most of the papers [4][5] has been presented using this dataset. Hence in this current work used UNSW-NB15 dataset instead of KDD99 dataset. Since dataset is labelled one so we implemented using various machine leaning algorithms.

UNSW (University of south wales)-NB15 [16] dataset contains raw network packet which was generated by IXIA perfect storm tool in the Cyber range lab of the Australian Centre for Cyber Security (ACCS) for producing hybrid of real modern normal activities and synthetic contemporary attack behaviours. This dataset was captured by tcpdump tool. This tool helped in capturing 100GB of data. A total of 49 attributes determining the features of connections are present for each data instances. This dataset include numeric, nominal and timestamp attributes. Consists of 9 types of attacks namely Analysis, Fuzzers, Backdoors, Daniel of Service (DOS), Exploit, Generic, Reconnaissance, Shellcode, and Worms. This dataset represented as modern synthesized attack of network and hybrid of computer network [5]. Since, labelled dataset we used for implementation of deep learning algorithms for network anomaly detection in supervised learning method. For our work, the UNSW-NB15 dataset contains 257673 information occurrences. The complete classes of this dataset are 10 classes: one is for a typical normal attack (93000 occurrences) and nine classes of a typical anomalous network data (assaults classes). The attacks included were backdoor (2329 instances), analysis (2677 occurrences), fuzzers (24246 cases), shellcode (1511), reconnaissance (13987 occasions), exploits (44525 cases), DoS (16353 cases), worms (174 cases), and generic (58871 cases).

B. Proposed Framework

Current paper focus is on a Deep learning model using modified support vector machine which includes both supervised and unsupervised learning models. System architecture is one which defines the Conceptual model in different structures and multiple views of the system. Fig. 1 shows architectural design of proposed system of the project. Above architecture clearly explains about how the components of the system communicate among themselves starting from pre-processing of data. This proposed framework is able to finding out the anomalies present in the computer networks data acquisition. Our System architecture flow is like starting from Network Data acquisition, here by using the wire shark, tcpdump tool etc. we can capture the network packets or network flow records with features.

This model gives clear picture of huge amount of network data capture and pre-processing of data to remove the noisy data presented in it. During pre-processing step we spit the dataset into a training and testing dataset. After this stem it should be stored into HDF5 (Hierarchical data format) because the Window system contains 4GB of Ram. Train dataset to detect the anomalies present in the dataset using appropriate supervised or unsupervised learning algorithms. Apply the deep leaning techniques which are helpful for finding anomalies for any of new data occurred in the computer network data acquisition. After this data acquisition suitable machine learning algorithm must be applied to compute effectivity and potentiality of the model, here we applied various machine leaning algorithms like support vector machine, Kmeans and Kohonen SOM etc. Metrics like accuracy, precision and Hit Rate will be calculated for the proposed model. This system architecture focuses 3 parts such as Real

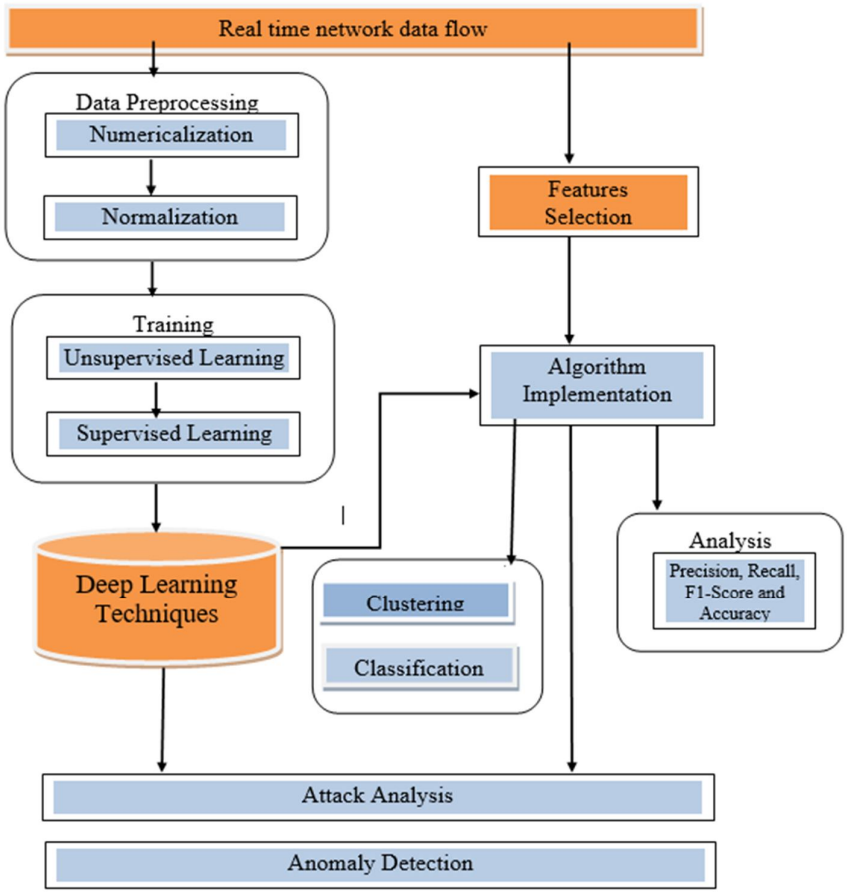

Figure .1 Block diagram of proposed system

time network flow data, Deep leaning techniques, and modules for detecting anomalies and feature selection modules.

C. Algorithms

Anomaly detection Algorithm
Nomenclature
Nd= Shuffled Network Traffic database.
Fs = Feature Selection from Network Traffic Database.
Ad = Anomaly Detection has to be predicted.
Bp = Back-Propagation Method.
Lb = Attack or not Label
Na = Number of Labelled Attack Category.
Td = Training Dataset.
Ts = Testing Dataset.
Vs = Validation Dataset.
Begin

Step 1: Select the Nd from the Network Data acquisition.
Step 2: Extract the Fs from the Nd Network traffic database.
Step 3: Split the Dataset into Td, Ts and Vs.
Step 4: Training Dataset has to be fed for training in deep neural network.
Step 5: Apply Back-Propagation Method for error-trial Purpose.
Step 6: Compute Td, Ts and Vs.
If Lb = Na

Ad Anomaly detection has to be predicted
Else
Carryout Multiple iteration Bp to detect the anomalies. End

IV. Results And Discussion

In this research, deep leaning frameworks played an important role in performing this experiment, this experiment is carried on personal notebook HP, which has a configuration of Intel Core i5- 4300U CPU @1.90 GHz, and 8 GB of memory. Significance of the task results are analyzed in this area. In this project we have picked 4 records of UNSW-NB15 dataset which are in .csv configuration has an all-out size of 570MB of data we have considered for the analysis. There will be 31.5 MB of training sample data and 15MB of testing data samples are taken for the deep neural network training model. Major thing in this project is feature extraction technique, choosing features from UNSW-NB15 dataset is important step. Initially we are choosing 5000 samples for training of model and chosen some of the features like Time, size of the packet and Location etc. Important features of the dataset are shown in below table 1. Gives details of training samples among total samples of given dataset.

TABLE I. TYPES OF ATTACKS

SI. No	**Type of attack**	**Training samples**	**Total Samples**
0.	Analysis	473	526
1.	Backdoor	480	534
2.	Dos	1050	1167
3.	Exploits	4868	5409
4.	Fuzzers	4545	5051
5.	Generic	5000	7522
6.	Normal	31000	677719
7.	Reconnaissance	1583	1759
8.	Shellcode	200	223
9.	Worm	20	24

As shown in below figures Fig2. Tells about feature importance of Random forest algorithm and Fig3. Tells about feature importance of neural network of attack analysis of total samples of UNSW-NB15 and all training data of the dataset. X axis tells about attack categories of UNSW-NB15 dataset and Y axis label explains about number of samples taken for attack analysis. These two graph explains efficiency of all the 9 categorical attacks of UNSW-NB 15 dataset.

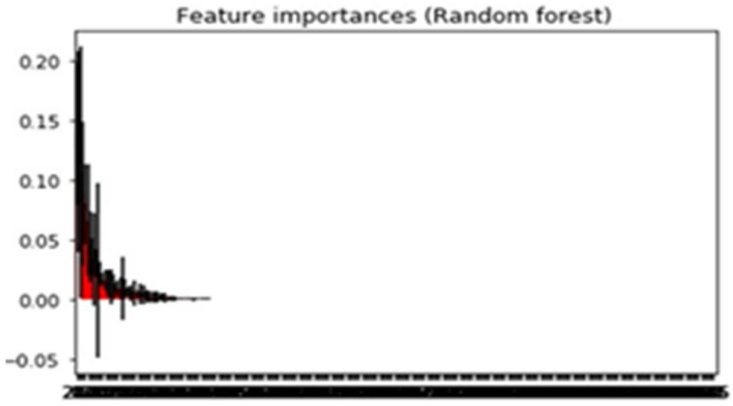

Figure 2. Feature importance of Random Forest

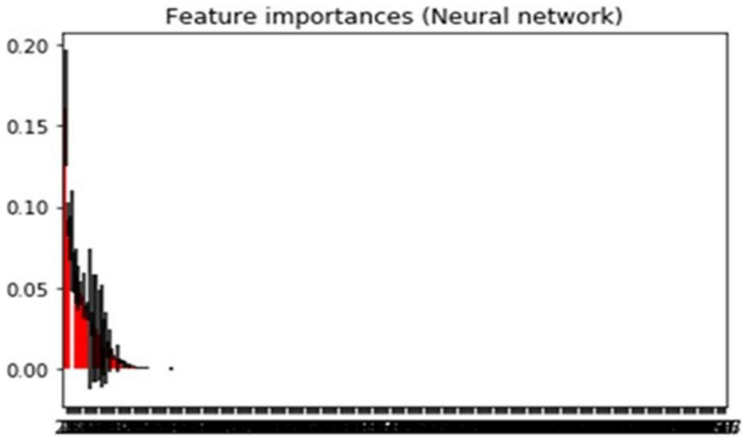

Figure 3. Feature importance of neural network

The below graph Figure 4. Shows feature importance of kohonen neural networks, SOM mapping steps begins from instating the weight vectors. SOM is a type of artificial neural network that is trained using unsupervised learning method to produce low dimensional discretized depiction of the input space of the training samples, called map. Figure 4. Explains anomalies of dependent and independent features of UNSW dataset w.r.t attack categories. Dark black line indicates winning vector of Kohonen SOM model.

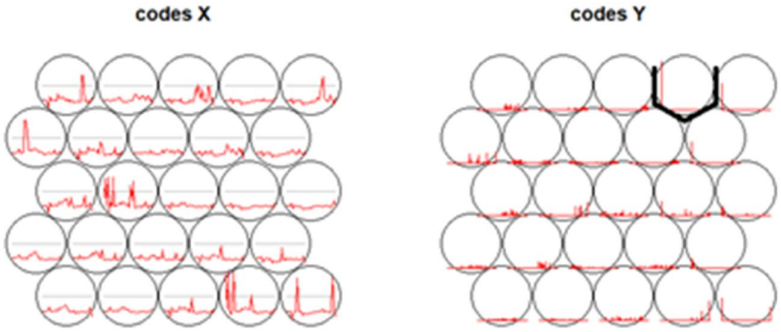

Figure 4. Feature importance of kohonen SOM

Fig5. Shows Anomalies from Kohonen self-organizing maps where Mean distance to the closest unit in Y axis, black line is matrix 1 is from which is based on independent variables and matrix 2 is from dependent variable data contained information about attack category.

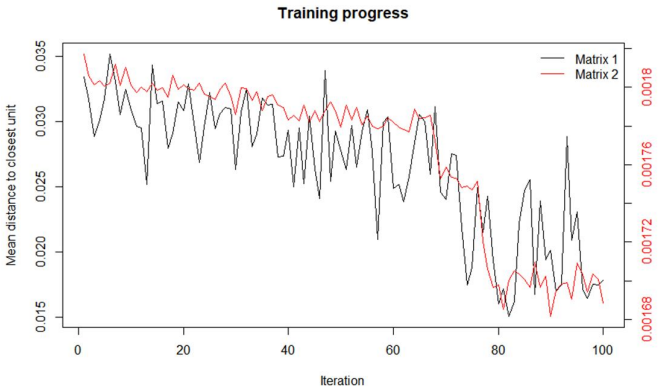

Figure 5. Anomalies from Kohonen Network

V. Conclusion

In this paper, we have proposed a Kohonen self-organizing maps and deep neural network approach for anomaly detection of computer network. SOM in Neural networks are trained using unsupervised learning is used to produce low dimensionality data. Also supervised learning is performed using support vector machine algorithm. The proposed model of Random forest yields the result of Accuracy 91.89%, which took time of 16 minutes to train and 8 seconds to classify the anomalies in the feature importance graph. Along with random forest, deep neural network model took a time 43 minutes and 3 seconds to train and classify the output that shows improvements in different dataset. We analyzed that proposed model has got more efficacy than the existing model for finding anomalies in computer network. We have presented a deep learning method anomaly detection of computer networks using various algorithms. And observations are effective and robust.

Acknowledgment

I take this opportunity to express my profound gratitude and deep regards to Head of Department and Professor Dr. Vijaya Kumar B. P. for his valuable information, exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

References

[1] Chuanlong yin , yuefei zhu, jinlong fei, and xinzheng he, “A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks”

[2] Ying Zhu.”Automatic Detection of Anomalies in Blood Glucose Using a Machine Learning Approac” 2011.

[3] 1mukrimah nawir, 2amiza amir, 3naimah yaakob, 4ong bi lynn. “Multi-classification of unsw-nb15 dataset for network anomaly detection system”, 2018.

[4] V. R. Balasaraswathi1, Ludovic Journaux and M. Sugumaran1. Benchmark Datasets for Network Intrusion Detection: A Review 2018.

[5] Monowar H. Bhuyan, D. K. Bhattacharya and J.K. Kalita, “Network Anomaly Detection:Methods, Systems and Tools” *IEEE Communications Surveys & Tutorials, VOL.16, NO. 1, First Quarter 2014.*

[6] Nguyen Thanh Van, Tran Ngoc Thinh, Le Thanh Sacg “An anomaly-based Network Intrusion Detection System using Deep Learning” *2017 International Conference on System Science and Engineering (ICSSE).*

[7] J.Schmidhuber, ‘‘Deep learning in neural networks: An overview,’’Neural network, vol. 61, pp. 85–117, Jan. 2015.

[8] Taeshik Shon’, Yongdue Kim, Cheolwon Lee’, and Jongsub Moon‘, “A Machine Learning Framework for Network Anomaly Detection using SVM and GA”2005.

[9] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita “Network Anomaly Detection: Methods, Systems and Tools” ieee communications surveys & tutorials, vol. 16, no. 1, first quarter 2014.

[10] Shan Ding and Genying Wang “Research on Intrusion Detection technology based on Deep Learning” 2017 3rd IEEE International Conference on Computer and Communication(ICCC).

[11] Phai Vu Dinh, Tran Nguyen Ngo, Nathan Shone, Aine MacDermott and Qi Shi “Deep- Learning Combined with de-noising data for Network Intrusion detection” 2017 21st Asia Pacific Symposium on Intelligent and Evolutionalry Sytems(IES).

[12] Lingli Lin, Shangping Zhong , Cunmin Jia and Kaizhi Chen “Insider Threat Detection Based on Deep Belief Network feature Representation” 2017 International Conference on Green Informatics(ICGI).

[13] Raghavendra Chalapathy, Aditya Krishna Menon and Sanjay Chawla “Anomaly Detection using One-Neural Networks”.

[14] Irfan A Siddavatam, S. Satish, W. Mahesh and Faruk Kazi “An Ensemble learning for anomaly identification in SCADA System” 2017 7th International Conference on Power Systems (ICPS)

[15] B.P Vijaya Kumar and P Venkatarm “Prediction-based location management using multilayer neural networks”J. Indian Inst. Sci., 2002, 82, 7–PRE

[16] Tharmini Janarthanan, Shahrzad Zargari."Feature selection in UNSW-NB15 and KDDCUP'99 datasets", 2017 IEEE 26thInternational Symposium on Industrial Electronics (ISIE), 2017

[17] B P Vijay Kumar, P Venkataram “Reliable multicast routing in mobile networks: a neuralnetwork approach” IEE Proceedings-Communications 2003/10/1.

[18] Vijay K Narayanan, Alok Kirpal & Nikos Karampatziakis “Anomaly Detection – Using Machine Learning to Detect Abnormalities in Time Series Data”.