IRJEAS
NURTURING RESEARCH

INTERNATIONAL RESEARCH JOURNAL
OF
ENGINEERING & APPLIED SCIENCES

ISSN: 2322-0821(O) VOLUME 6 ISSUE 1
ISSN: 2394-9910(P) Jan 2018 - Mar 2018

www.irjeas.org

# CLASSIFICATION OF CYBER ATTACK USING MACHINE LEARNING TECHNIQUE AT MICROSOFT AZURE CLOUD

**Rahul Chourasiya[1], Vaibhav Patel[2], Anurag Shrivastava[3]**

[1]*PG Scholar, Dept. of* Computer Science Engineering, NIRT*, Bhopal.*
[2]*Asst. Professor, Dept. of* Computer Science Engineering, NIRT*, Bhopal.*
[3]*Associate Professor & Head, Dept. of* Computer Science Engineering, NIRT*, Bhopal.*

***Abstract— Security of valuable information is always a very essential issue for modern digital world. Intrusion Detection System (IDS) and many security techniques is widely used against cyber attacks. Data mining and machine learning methods have also been used by researchers to obtain high detection rate and low false alarm rate. Proposed work aims to design and development of an approach for improve cyber attack detection system using cloud. Uses of cloud computing is increases very progressively. Using traditional ML Techniques do not support well processing of large datasets, so new approaches and platforms are needed. This paper proposes that cloud based machine learning technique can be used in order to classify attack into a cloud based machine learning platform. The work proposes a attack classification framework using NSL KDD Cup99 dataset. The classifier is build which is based on 'Multiclass Decision Forest' Machine Learning Algorithm and is deployed on Microsoft's Azure Machine Learning (Azure ML) platform. Azure ML is public cloud platform. The results obtained by proposed model are evaluated in terms of accuracy and the comparison is done with benchmarks provided by competition administrators. The results obtained are promising and the paper also directs the future research work in the field.***

***Keywords— Cloud Computing, Cyber Attack, IDS, Classification, Machine Learning, Microsoft Azure Cloud***

## 1. INTRODUCTION

### 1.1 Machine Learning and IDS :

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. Machine learning techniques have ability to implement a system that can learn from data. For example, a machine learning system could be trained on incoming packets to learn to distinguish between intrusive and normal packet. After learning, it can then be used to classify new incoming packets into intrusive and normal packets.[23] In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents task. A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. Machine learning often included in the category of predictive analytics as it helps to predict the future analysis.

Intrusion Detection System (IDS) is an active process or device that analyzes system and network activity for unauthorized activity [2]. An ID is hardware or software or a combination of both which is used to monitor a system or network of systems against any malicious or unauthorized activities [2]. Intrusion Detection Systems (IDSs) are used to improve network security. An ID improves the security of the network by identifying, assessing, and reporting unauthorized network activities. IDS are categorized into two classes: network-based and host-based. Network based Intrusion Detection Systems analyses network packets retrieved from the network. Host-based Intrusion Detection System analyses system calls generated by individual hosts [2].The data flows through a network is very large and it is difficult to analyze and detect the attacks using traditional methods. Today we have number of Machine learning techniques available which are very useful for analyzing the data and detecting the attacks. In this paper we have used various machine learning techniques for network intrusion detection [2].

### 1.2 IDS in Cloud :

Intrusion detection system plays an important role in the security and perseverance of active defense system against intruder hostile attacks for any business and IT organization. IDS implementation in cloud computing requires an efficient, scalable and virtualization-based approach. In cloud computing, user data and application is hosted on cloud service provider's remote servers and cloud user has a limited control over its data and resources. In such case, the administration of IDS in cloud becomes the responsibility of cloud provider. Although the administrator of cloud IDS should be the user and not the provider of cloud services.[25]

**1.3 Microsoft Azure Cloud Computing Environment for Machine learning:** Microsoft's Azure Machine Learning (Azure ML) [3] is a cloud service that enables execution of machine learning process. Microsoft Azure is a public cloud platform. The benefits of using public cloud computing platform (Azure ML) includes: handling big data and access from anywhere in the world. The process of Azure ML is shown in Figure – 1, which is same as that of basic process of ML. Azure ML provides a graphical tool for managing the ML process, a set of data pre-processing modules, a set of machine learning algorithms, and an API to launch a model to applications. ML Studio is a graphical tool that is used to control the process from beginning to end i.e. from data pre-processing to run experiments using a machine learning algorithm, and test the resulting model. ML Studio also helps its users deploy that model on real cloud.

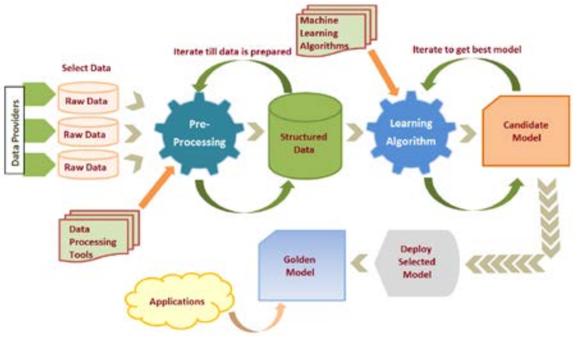


**Figure 1: Machine Learning Process**

The need of cloud platforms to classify NSL KDD data is established in next section. The rest of the paper organized as: Section 2 briefly surveys the need of cloud platforms for IDS in intrusion dataset. The work proposed is presented in section 3.Experimental setup and result analysis is shown in section 4 and paper is concluded in section 5.

## 2. Literature Review (Need of Cloud Platforms for IDS)

Traditional Intrusion detection system using data mining and machine learning techniques are work on information system they are not working on cloud environment. Here give some literature about Intrusion detection system and using cloud for classification with machine learning algorithms. Multiple choices of cloud computing models are available for different work load management, performance and computational requirements. The popular statistical tools and environments like Octave, R and Python are now embedded in the cloud as well [5]. The important findings of work [6] indicate the area of customer retention received most research attention.

**Machine Learning on Cloud environment for Fast Prediction in Big Data:** As the data is growing at faster rate and becoming "Big Data", the computation speed for prediction and other operations is inevitable. This paper [7] focused on the specific problem of classification of network intrusion traffic which is a Big Data.

Authors [3] worked on IDS for web proxy, taking inspiration from Intrusion Detection Systems that make use of machine learning capabilities to improve anomaly detection accuracy, this paper proposes that cloud based machine learning can be used in order to detect and classify web proxy usage by capturing packet data and feeding it into a cloud based machine learning web service.

In this paper, [22] authors examine different machine learning techniques that have been proposed for detecting intrusion by focusing on the hybrid classifier algorithms. The objective is to determine their strengths and weaknesses. From the comparison, authors hope to identify the gap for developing an efficient intrusion detection system that is yet to be researched.

Authors [23] said about the cloud based attack system.

Authors add new valued feature to the cloud-based websites and at the same time introduces new threats for such services. DDoS attack is one such serious threat. Covariance matrix approach is used in this article to detect such attacks. The results were encouraging, according to confusion matrix and ROC descriptors.

Authors [25] proposed cloud IDS handles large flow of data packets, analyze them and generate reports efficiently by integrating knowledge and behavior analysis to detect intrusions.

Authors [26] proposed anomaly Intrusion Detection System using machine learning approach for virtual machines on cloud computing. In this work feature selection over events from Virtual Machine Monitor to detect anomaly in parallel to training the system so it will learn new threats and update the model. The Proposed experiment has been carried out on NSL-KDD'99 datasets using Naïve Bayes Tree (NB Tree) Classifier and hybrid approach of NB Tree and Random Forest.

## 3. PROPOSED FRAMEWORK FOR CLASSIFICATION:

The Proposed Framework which employs simple machine learning (ML) model with little change. The input KDDcup99 dataset is suitably processed and converted into a suitable format. The machine learning algorithms are iteratively applied in the next step, and candidate model is determined. These ML algorithms typically apply some statistical analysis like regression or more complex approaches like decision forest to the data. Here in the proposed framework, the ensemble methods [12] are also applied to the model for better accuracy. At last the model is deployed and tested on test data the snapshot of actual model build using specified steps, at Microsoft Azure ML platform, is shown in Figure – 2.

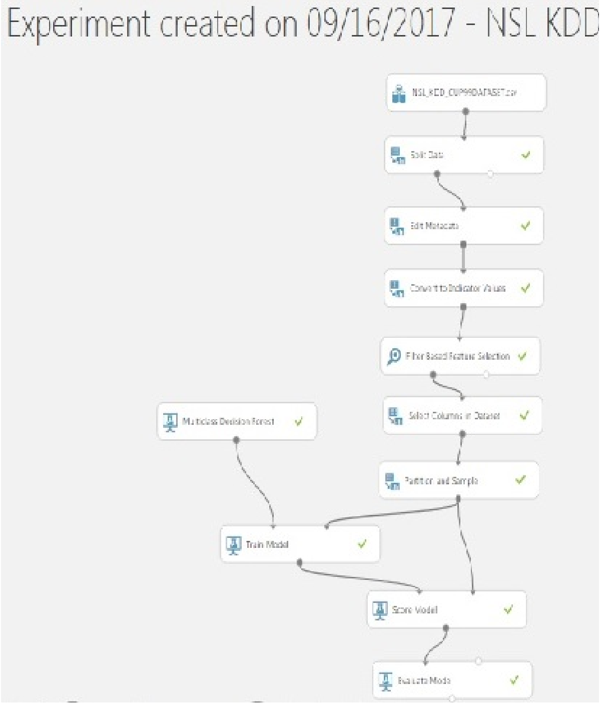

Experiment created on 09/16/2017 - NSL KDD
NSL_KDD_CUP99DATASET.csv
Split Data
Edit Metadata
Convert to Indicator Values
Filter Based Feature Selection
Multiclass Decision Forest
Select Columns in Dataset
Partition and Sample
Train Model
Score Model
Evaluate Model

Figure 2: Model built using Azure ML

## 4. Simulation Environment Setup and Result Analysis:

Azure ML provides ML studio, a graphical tool that can be used to control the process from beginning to end. It includes: a set of data pre-processing modules; a set of machine learning algorithms; An Azure ML API to access model deployed on Azure. ML Studio allows a user to import datasets and data pre-processing methods.

**4.1 NSL KDD CUP 99 DATASET:** In Earlier days the researcher focused on DARPA dataset for analyzing intrusion detection [13]. It consists of seven weeks of training and also two weeks of testing raw tcpdump data. The main drawback is its packet loss. The refined version of DARPA dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset [11]. Which consist on 5 million single connections for training records and 2 million connections for testing Due to its huge size the researcher used on 10% of dataset to analysis intrusion accuracy which affects the performance of the system, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [27] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset is 1. No redundant records in the train set, so the classifier will not produce any biased result 2. No duplicate record in the test set which have better reduction rates. 3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set. The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test dataset i.e. not available in the training datasets. The details of attack categories and specific types are shown in Table1. According to Table1, there are four attack categories in NSL KDD99 dataset:

(1) Probing: Scan networks to gather deeper information
(2) DoS: Denial of service
(3) U2R: Illegal access to gain super user privileges
(4) R2L: Illegal access from a remote machine.

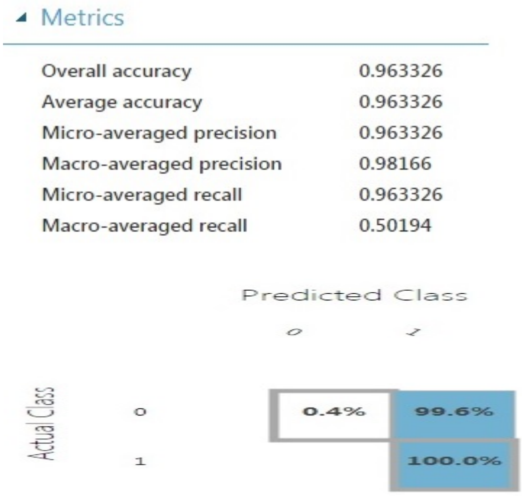
**4.2 Execution of Implemented Work (Experiment Steps):** The experimental steps that are and represented in Figure–2, are explained below:

1. Create New Resource: Machine Learning Analytics solution.
2. Import/Upload the dataset.
3. Pre-process the dataset. Data pre-processing can also be done using modules written in R or Python.
4. Randomly split and partition the data into 70% training and 30% testing, using the ‘Split Data’ module.
5. Identify categorical attributes and cast them into categorical features using the ‘Edit Metadata’ module.
6. Convert to Indicator Values module to convert columns that contain categorical values which can more easily be used as features.
7. Select Columns in Dataset those are relevant
8. Apply Ensemble Method
9. Apply Machine Learning Algorithm to Train the model.
10. Now Score and Evaluate the Model. The ‘Evaluate model’ also visualizes the results through confusion matrix.

**4.3 Experimental Results Analysis and Discussion**: The experiment is evaluated on a simple multi-class decision forest classification accuracy parameter. Accuracy is defined as the number of correctly classified instances divided by the total number of instances:

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{Number of Instances}}$$

The results obtained using the benchmark code by setting the multicast decision forest model got the accuracy of 0.9633 in experiment, while the benchmark results given by competition administrators with is 0.50241. Here we have performed experiment at cloud platform with Multiclass decision forest techniques with an ensemble method. The evaluation results are inferred from confusion matrix shown in Figure – 3. A confusion matrix also known as error matrix and is used to describe the performance of a classifier (classification model). The overall accuracy obtained with our simulation is 0.9633, which is higher than the benchmark provided. The comparison of proposed model is done with benchmark provided by administrators and competition’s winning results.



Metrics

| | |
|---|---|
| Overall accuracy | 0.963326 |
| Average accuracy | 0.963326 |
| Micro-averaged precision | 0.963326 |
| Macro-averaged precision | 0.98166 |
| Micro-averaged recall | 0.963326 |
| Macro-averaged recall | 0.50194 |

| Actual Class \ Predicted Class | 0 | 1 |
|---|---|---|
| 0 | 0.4% | 99.6% |
| 1 | | 100.0% |

**Figure 3: Confusion Matrix with Multicast Decision forest**

The comparison for accuracy obtained, is shown in Figure–4.

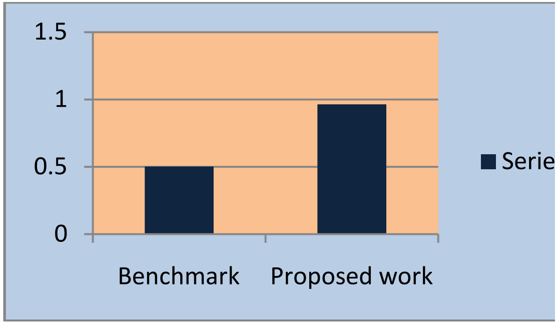


**Figure 4: Comparison for Accuracy**

## 5. Conclusion & Future Work

In this paper, Machine Learning technique has been proposed in terms of accuracy and detection rate for four categories of attack under different percentage of normal data. The purpose of this proposed method efficiently classify abnormal and normal data by using very large data set and detect intrusions even in large datasets with short training and testing times. With proposed method we get high accuracy for many categories of attacks and detection rate with low false alarm. In this paper, we proposed an Azure ML based model for attack classification. The model used Multicast Decision forest algorithm to train the classifier. The evaluation results show that the proposed classifier performs better in terms of accuracy. We have performed experiment with multicast decision forest and an ensemble method. Our experiments showed the better accuracy than benchmark .The proposed research can provide potential approach for training and testing of big data for addressing multi-class classification problems. So, further research will evaluate the framework with different ML algorithms. In future the model can be optimized to handle imbalanced datasets from various sources and domains. Also, the model can be modified for applying on Hadoop MapReduce [11] platform.

## References

[1] Pine II, B.J. and Gilmore, J.H. 1999. The Experience Economy. Boston: Harvard Business School Press.

[2] Ch.Ambedkar, V. Kishore Babu, "©ARC Page 25 Detection of Probe Attacks Using Machine Learning Techniques" International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 2, Issue 3, March 2015, PP 25-29 ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)

[3] Shane Miller, Kevin Curran, Tom Lunney " Cloud-based machine learning for the detection of anonymous web proxies" ISSC 2016.

[4] David Chappell, "introducing azure machine learning: a guide for technical professionals", Sponsored by Microsoft Corporation, 2015 Chappell & Associates.

[5] https://portal.azure.com

[6] Daniel Pop, "Machine Learning and Cloud Computing Survey of Distributed and SaaS Solutions", https://www.researchgate.net/publication/257068169.

[7] E.W.T. Ngai ,, Li Xiu, D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications 36 (2009) 2592–2602, Elsevier

[8] Suthaharan, S., "Big data classification: Problems and challenges in network intrusion prediction with machine learning" Performance Evaluation Review, 41(4), 70-73, ACM 2014.

[9] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze "A Novel Intrusion Detection Method Based on Support Vector Machines" IEEE 2010.

[10] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest", R News, ISSN 1609-3631, Vol. 2/3, December 2002.

[11]https://www.MulticlassDecisionForest.html

[12]Apache Hadoop Website http://hadoop.apache.org/

[13] J. a. H. Friedman, Trevor and Tibshirani, Robert, The elements of statistical learning vol.1: Springer series in statistics Springer, Berlin, 2001.

[14] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Gorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", 2009 IEEE

[15]YU-XIN MENG "The Practice on Using Machine Learning For Network Anomaly Intrusion Detection" 2011 IEEE

[16] Chi Cheng, Wee Peng Tay and Guang-Bin Huang "Extreme Learning Machines for Intrusion Detection" - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[17]Solane Duquea*, Dr.Mohd. Nizam bin Omarb Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS) www.sciencedirect.com, Elsevier 2015.

[18]Naeem Seliya , Taghi M. Khoshgoftaar "Active Learning with Neural Networks for Intrusion Detection" IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10/$26.00 ©2010 IEEE

[19]Kamarularifin Abd Jalill, Mohamad Noorman Masrek "Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion" 201O

International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[20] Shingo Mabu, Member, IEEE, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa, Member, IEEE "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming" IEEE, JANUARY 2011

[21] Liu Hui, CAO Yonghui "Research Intrusion Detection Techniques from the Perspective of Machine Learning" - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[22] 1sundus juma, 1zaiton muda, 1m.a. mohamed, 2warusia yassin "machine learning techniques for intrusion detection system: a review" Journal of Theoretical and Applied Information Technology 28th February 2015. Vol.72 No.3

[23] Anku Jaiswal, Chidananda Murthy P, Madhu BR "Prevent DDOS Attack in Cloud Using Machine Learning" Volume 6, Issue 6, June 2016 ISSN: 2277 128X

[24]Abdulaziz Aborujilah1 and Shahrulniza Musa2 "Cloud-Based DDoS HTTP Attack Detection Using Covariance Matrix Approach" Hindawi Journal of Computer Networks and Communications
Volume 2017, Article ID 7674594, 8 pages
https://doi.org/10.1155/2017/7674594

[25] Ms. Parag K. Shelke, Ms. Sneha Sontakke, Dr. A. D. Gawande "Intrusion Detection System for Cloud Computing" International Journal of Scientific & Technology Research Volume 1, Issue 4, May 2012 ISSN 2277-8616

[26] Amjad Hussain Bhat1, Sabyasachi Patra2, Dr. Debasish Jena3 "Machine Learning Approach for Intrusion Detection on Cloud Virtual Machines" Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com Volume 2, Issue 6, June 2013.

[27]Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze "A Novel Intrusion Detection Method Based on Support Vector Machines" IEEE 2010.