IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107003 |

A Cloud-Based Intrusion Detection Model Using Principal Component Analysis and Random Forest Approach

Oyefia George Emakpo, Taylor Onate Egerton, Nwaiabu Nuka Dumle

Msc Student, Dept. of Computer Science, Rivers State University of Science and Technology, Port Harcourt, Rivers State, Nigeria

Senior Lecturer, Dept. of Computer Science, Rivers State University of Science and Technology, Port Harcourt, Rivers State, Nigeria

Senior Lecturer, Dept. of Computer Science, Rivers State University of Science and Technology, Port Harcourt, Rivers State, Nigeria

ABSTRACT: Technological advancement in wireless communication has resulted in an array of security threats prevalent over the cloud. Intrusion detection system (IDS) facilitates the identification of any potential network attacks on a cloud. There are existing systems such as the Support Vector Machine (SVM) and Naïve Bayes which has been implemented to identify network attacks on a cloud. The objective of this study was to improve on the accuracy and the error rate of these existing systems. A constructive research methodology and a waterfall model design methodology was adopted for this study, a model was developed using the Python programming high level language, and a principal component analysis (PCA) technique was employed to reduce the dimensionality of the input KDD dataset, while the random forest algorithm was used to train the model for the identification of attacks. The findings showed that the model exhibits better accuracy and error rate in comparison to existing systems such as the Support Vector Machine (SVM) and Naïve Bayes, the proposed approach yielded results characterized by an accuracy rate of 99.21%, and an error rate of 0.79%, as against Support Vector Machine (SVM) of an accuracy rate of 84.34%, and an error rate of 2.67 %, Naïve Bayes of an accuracy rate of 80.85% and an error rate of 3.49%.

KEYWORDS: Intrusion Detection System (IDS), SVM; KDD dataset, Principal Component Analysis (PCA), Random Forest

I. INTRODUCTION

Cloud computing is a web based computing where virtual shared servers allocates programs, frameworks, platforms, gadgets ,assets and hosting to a client as a service based on pay-as you-use. All the information that a digitized framework has got to offer is given as a service within the cloud computing model. Clients can get to these administrations accessible on the "internet cloud" without having any previous know-how on managing the assets included. Cloud clients don't own the physical infrastructure; they lease the utilization from a third- party supplier. They utilize resources as a service and pay for only the resource that they utilize. What they as it were require may be an individual computer and web connection. Cloud computing has revolutionized the IT world with its administrations providing framework, less maintenance cost, information & services availability assurance, rapid openness and scalability. Cloud computing has three fundamental layers i.e. system layer (which may be a virtual machine abstraction of a server), the platform layer (a virtualized working framework of a server) and application layer (that incorporates web applications). Hardware layer isn't included because it does not straightforwardly offered to clients. Cloud computing has three service models known as Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (SaaS) models. PaaS model encourages clients by giving platforms on which applications can be created and run. IaaS deliver services to clients by keeping up huge infrastructures like facilitating servers, overseeing systems and other assets for clients. SaaS demonstrate makes client stress free of installing and running computer service on its own machines. At present, Salesforce.com, Google, and Amazon are among the prominent cloud service providers that offer their services in the areas of storage, application, and computation through a pay-as-you-go basis. The unavailability of data, applications, and services can be inflicted through Denial of Service (DOS) or Distributed Denial of Service (DDOS) attacks, which can render both cloud service providers and users incapable of delivering or receiving cloud services. To mitigate these types of attacks, the deployment of an Intrusion Detection

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107003 |

System (IDS) serves as a robust defensive mechanism. In the aspect of computer security, Intrusion Detection Systems (IDSs) emerge as a crucial technology for detecting and mitigating various types of cyber threats. More specifically, IDSs take on different forms, namely host-based, network-based and distributed IDSs. (Chai and Bigelow., 2022.).

II. RELATED WORK

Geramiraz *et al*. (2017), propose an integrated machine learning calculation based on K-Means clustering and the Naive Bayes Classifier (NBC) named KMC+NBC to maximize discovery and accuracy whereas minimizing wrong alarm. K-Means clustering has been connected to the labeling handle. All the information are collected in their comparing clusters as typical and forceful concurring to their behavior with K-Means, whereas the Naive Bayes Classifier is utilized to recategorize the misclassified information into the proper categories. Execution evaluations of KMC+NBC and NBC were made on the ISCX-2012 dataset. Agreeing to the results obtained, KMC+NBC expanded the accuracy and discovery rate p to 97% and 98.8%, ndividually, whereas diminishing the wrong caution to 2.2%. This ponder can be extended to incorporate include feature selection methods before machine classification.

Singh *et al*. (2021), proposed a novel calculation based on Profound learning procedure to secure the Remote systems from the assaults and detect any such activity. This proposed calculation uses the Customized Rotation Forest calculation for the point of selecting features. The classification of diverse attacks is carried out by Gated Repetitive Units (GRU). The displayed show was connected on NSL-KDD dataset and a 97.32% and 97.47% accuracy rate were gotten for twofold and multiclass categorization, separately. According to comes about, the proposed WIDS can be utilized for recognizing assaults in real-time systems. Be that as it may, this study can be moved forward by applying the proposed demonstrate on distinctive datasets.

III. PROPOSED ALGORITHM

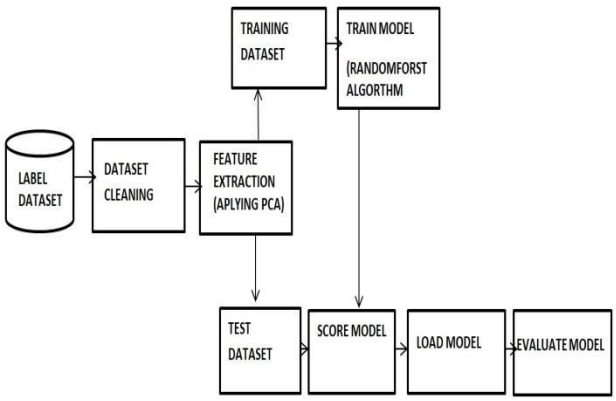

Figure 1: The Proposed architecture of the model

Principal Component Analysis

The principal component analysis is the technique that is used, especially for the reduction of the dimension of the given dataset. The principal component analysis is one of the most efficient and an accurate method for reducing the dimensions of data, and it provides the desired results. This method reduces the aspects of the given dataset into a desired number of attributes called principal components.

This method takes all the input as the dataset, which is having a high number of attributes so as the dimension of the dataset is very high. This method reduces the size of the dataset by taking the data points on the same axis. The data

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107003 |

points are shifted on a single axis, and the principal components are carried out. The PCA can be performed using the following steps:

1. Take the dataset with all dimensions d.
2. Calculate the mean vector for each dimension d.
3. Calculate the covariance matrix for the whole dataset.
4. Calculate the eigen vectors (e1, e2, e 3 … .ed), and eigen values (v1,, v2, v3,….vd).
5. Perform sorting of eigenvalue in decreasing order and select n eigenvector with the highest eigenvalues to get a matrix of d*n= M.
6. By using this M form a new sample space.
7. The obtained sample spaces are the principal components.

Random Forest Algorithm

Random Forest is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classifier.

IV. SIMULATION RESULTS

The experiment carried out for the proposed approach uses the KDD dataset, and the results obtained were satisfying. The following configurations are used for performing our analysis

I. In Hardware: 4 GB RAM, 140 Gb SSD Hard disk, Intel core i3 and intel motherboard. ¬
II. In Software: 64-bit windows 10 and Python 3.8. ¬ Python packages like NumPy, flask and KerasLibrary
III. Data set: KDD dataset.

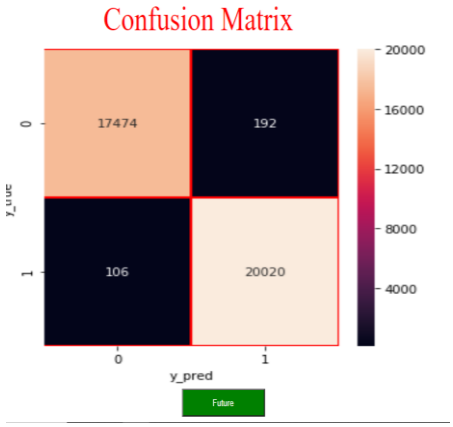

Figure 2: A Confusion matrix analysis of the result

From the confusion matrix predicted by the model, The True Negative is given as 17474, The false Positive is given as 192 ,The false negative is given as 106 ,The True Positive is given as 20020.

i. **Accuracy** =TP + TN /Total actual Predicted
20020+17474/37792 = 99.21%
ii. **Error rate** =FP + FN /Total actual predicted
192 + 106 /37792 =0.79%

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 7, July 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1107003 |

Table 1: Tabular Comparison of result with other Classifiers

Method	Accuracy	Error Rate
SVM	83.34	2.67
Naïve Byte	80.85	3.49
PCA With Random Forest	99.21	0.79

The table given above gives a numerical representation of the obtained values from the experiment. The error rate found in our proposed approach is very low as of .21%. As well, the accuracy obtained is much higher than previous algorithms. Also, the time taken for the performance is less than other algorithms.

V. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed model performs better with the obtained values for Accuracy rate and error rate were 99.21% and 0.79%, respectively. The model under consideration has exhibited commendable performance in comparison to the previously utilized algorithms namely Support Vector Machine (SVM) and Naïve Baye. In future work, it is suggested to explore different machine learning techniques, to enhance accuracy, use a more current dataset for evaluating the model's efficiency. The rise of advanced technology has led to an increase the level of connectivity in the cloud, the prevailing computational capacity for machine learning is lacking and will prove inadequate in the upcoming times. Thus, the progression of neural network adaptation must be adopted.

REFERENCES

1. Chai, W. & Bigelow, S. J. (2022). Cloud computing. *Cloud Computing*.
2. Geramiraz, F., Memaripour, S. & Abbaspour, M. (2017). Adaptive anomaly based intrusion detection system using fuzzy controller. *Int. J. Network. Security*, *14*(6), 352–361.
3. Singh, N.B., Singh, M.M., Sarkar, A. & Mandal, K. (2021). 'A novel wide & deep transfer learning stacked GRU framework for network intrusion detection. *J. Inf. Secur. Appl*, *61*, 102899.
4. Great Learning Team. (2023). Random forest Algorithm in Machine learning | Great Learning. Great Learning Blog: Free Resources What Matters to Shape Your Career.
5. Jaadi, Z. (2021). A Step-by-Step Explanation of Principal Component Analysis (PCA). *Built In*.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details