International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 5 May 2023, pp: 856-859 www.ijaem.net ISSN: 2395-5252

Network Intrusion System Using PCA and Random Forest

Utkarsh, Abhishek kumar, Akash, Megha Shalot
Department of computer science and engineering, Raj Kumar Goel Institute of Technology, Ghaziabad U.P.

Indira Adak
(Asst. Professor), Department of computer science and engineering, Raj Kumar Goel Institute of Technology, Ghaziabad U.P.

Date of Submission: 10-05-2023 Date of Acceptance: 23-05-2023

ABSTRACT:
Due to the advancement of wireless communication, there are several online security risks. The intrusion detection system (IDS) assists in identifying system attacks and identifies attackers. In the past, several machine learning (ML) techniques have been applied to IDS in an effort to improve intruder detection outcomes and boost IDS accuracy. In this paper, a method for creating effective IDS that makes use of the random forest classification algorithm and principal component analysis (PCA) is proposed. Whereas the random forest will aid in classifying while the PCA will assist in organizing the dataset by lowering its dimensionality. Results show that the suggested approach performs more accurately and efficiently than compared to SVM, Naive Bayes, and Decision Tree methods.
Keywords: Principal Component Analysis,IDS,KDD, Random Forest,Machine Learning.

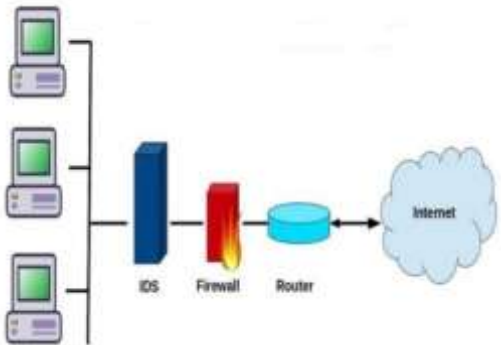
I. INTRODUCTION:

An "intrusion" occurs when someone enters a system without being invited and tampers with the data that is already there. This may potentially harm the hardware of any system.intrusion. The intrusion has turned into a keyword to defend the system from. This intrusion can be controlled or monitored inside any system with the help of the I DS. Despite the use of a variety of intrusion detection systems in the past, accuracy problems have been discovered with each method. Two concepts—detection rate and false alarm rate—are researched in order to gauge how accurate the system is. These two sentences should be drafted to minimize the possibility of false alarms and boost the system's detection rate. Therefore, the (IDS) is applied using both the PCA and random forest. It can be used for the following two different types of IDS:

• Systems that assess network traffic and intrusions over it are known as network intrusion detection systems (NIDS).
• System checks network access to system files in the case of host-based intrusion detection systems (HIDS). There is also an IDS subgroup. The most often used versions rely on signature and anomaly detection.

Signature-based: The system identified specific patterns used by malware in this technique.

These discovered patterns are known as signatures. This is effective at locating recent assaults, but it is useless at findingrecent assaults.



• Anomaly-based: This technique was developed especially for spotting unknown attacks. This system's ML is used to build the model.

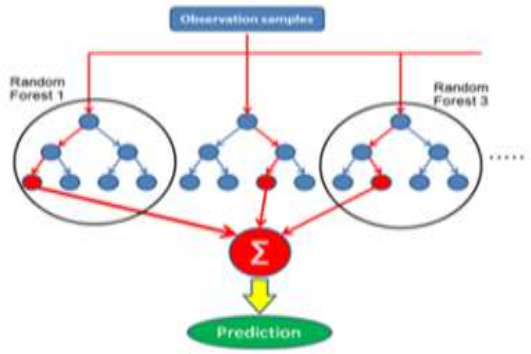
Random Forest

Decision tree algorithms are the foundation of the random forest-supervised machine learning technique. Several businesses, including banking and e-commerce, employ this method to predict behavior and outcomes. This page provides an

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 5 May 2023, pp: 856-859 www.ijaem.net ISSN: 2395-5252

explanation of how the random forest algorithm works. The essay will discuss the features of the algorithm and how it is applied in real-world scenarios. It also points out the advantages and disadvantages of this algorithm.

1. Select a few characteristics k as km from all m.
2. Apply split point from k characteristics to obtain node d.
3. To get the daughter nodes, use the best split.
4. Repetition of the first three stages should continue until you reach one node.
5. To make a forest, repeat steps 1 through 4 again.



PCA

The principal component analysis is the method used, which is particularly helpful for lowering a dataset's dimension. One of the most effective and precise techniques for reducing data dimensionality is principal component analysis, and it produces the desired outcomes. This process streamlines a's properties. a given dataset into the necessary number of primary components, or features.

The dataset created using this method, which has a huge number of characteristics and a large dimension, utilizes all of the input. This method reduces the size of the dataset by aligning the data points on the same axis. The primary operations are carried out while the data points are shifted along a single axis. There are several techniques to conduct a PCA:

1. Take into account the d-dimensional dataset.
2. Establish the mean vector for every d-dimensional variable.
3. Construct an overall dataset covariance matrix.
4. Calculate the eigenvalues (v1, v2, v3....yd) and the eigenvectors (e1, e2, e3...ed).
5. To get a matrix with d*n=M, sort the eigenvalues in decreasing order and select the n eigenvector with the highest eigenvalues.
6. Use this M to make a new test place.

II. LITERATURE REVIEW:

Machine learning (ML) is rapidly being used in the field of network intrusion detection to analyse data and create compelling models. This section of the article provides an overview of the research on the many ways machine learning can be applied to improve network intrusion system performance. An approach for building an IDS for the Internet of Things has been offered by the authors of this article, and it is based on categorizing traffic using a machine learning model. In this paper, we provide an intrusion detection approach that makes use of the Random Forest algorithm and Principal Component Analysis (PCA). In this case, the random forest will help with categorization while the PCA will help organise the dataset by reducing its dimensionality.

The KDD dataset will be used as the data set for the third International Knowledge Discovery and Data Mining Tools Competition.To develop it, 41 features from the DARPA offline were used. Detection of intrusions evaluation.

Attacks are categorized into four primary categories. Denial of Service (DoS), Probe, User to Root (U2R), and Remote to Local (R2L) are the four types of attacks.

Attacks in this category are designed to collect information for possible infiltration. DoS: Denial of Service (DoS) attacks stop a host or server from performing its normal functions by crashing it.R2L: Remote to Local (R2L), sometimes known as remote-to-local authentication, is a subset of this category that enables an attacker to run commands by getting around normal authentication.

Attacks from User to Root (U2R) fall into this category. In these attacks, an attacker impersonates the network's root user by obtaining login credentials from authorized users.

The test data does not come from the same probability distribution as the training data and does contain some unique attack types that were not present in the training data. This is somewhat comparable to the actual situation where inventive intrusion occurs. Due to the size of the total data set, training and test sets were built using 10% selections.

To give the dataset to the intrusion detection system, the authors enhanced its quality. They created a feature selection method based on fuzzy rules in order to improve the dataset. The KDD dataset was used, and the IDS results demonstrated dynamic growth.

Methodology.

This paper's methodology section describes the technique used to investigate the potential

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 5 May 2023, pp: 856-859 www.ijaem.net ISSN: 2395-5252

applications of machine learning to improve system performance. The methods employed in this study for gathering data and analyzing it are both included in the methodology.

Data Collection

The KDD Dataset was primarily used to gather the data for this investigation. We conducted a thorough search of several academic databases, including Google Scholar, IEEE Xplore, and the ACM Digital Library, using various phrase combinations, such as "machine learning," "KDD," and "optimization." Depending on how relevant they were to the topic of the study, 30 scholarly papers and case studies in total were selected for analysis.

One of the KDD data set's biggest flaws is the vast amount of redundant records because it forces learning algorithms to favor learning frequent records and prevents them from learning less frequent records, which are frequently more dangerous to networks, such as U2R and R2L attacks. Additionally, adding these repeated records to the test set will skew the evaluation results.

III. DATA ANALYSIS

Principal Component Analysis (PCA) and the Random Forest classification technique were used to analyse the data that was gathered. Whereas the random forest will aid in classifying while the PCA will assist in organizing the dataset by lowering its dimensionality.

The two approaches that make up the proposed system are random forest and principal component analysis. Principal component analysis reduces the dataset's dimension, which raises the possibility that the dataset has the appropriate attributes and raises the dataset's quality. The random forest method will then be used in the intruder detection process, which, when compared to SVM, delivers an improved detection rate and a lower false alarm rate.

Attribute compatibility has taken the place of the original attribute for the split node standard, coordination degree.

1. Attribute compatibility

The proposed algorithm's flowchart is as follows: CO(X D) = [Pr] [Se] | K] Let the modulus of the primary decision set be | Pr |, the modulus of the secondary decision set be | Se |, and attribute compatibility be specified as 2.

Here, X is the subset for non-empty C. When the mind is affected by the secondary set It is claimed to be in strict compatibility because it is obvious. They are at odds with one another: the first set and the second set. The expression brings the second set to a close.

CO(XD)=PR/K. Here, X is the subset for non-empty C. This illustrates just how compatible the second set is overall:

Improve the Base Classifier Algorithm:

Mark each condition attribute that is being used to iterate the data set in Step 1.

Step 1: In the primary and secondary sets, ascertain the modulus for each condition characteristic.

phase 3: Equation 1 is used to do all conditional attribute compatibility calculations in this phase. Use equation 2 if other features are found to be compatible.

Step 4: Select the split node that has the greatest compatibility for splitting the sample and delete the active tag.

Step 5: Up until the leaf node, keep picking the active attribute for splitting.

The base classifier is eventually generated in step six.

IV. LIMITATIONS

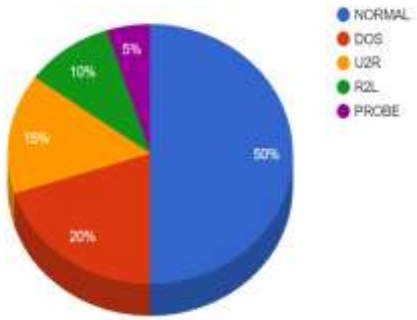
Online systems are susceptible to a variety of damaging actions. The fundamental problem with this topic is the information breach caused by system infiltration.

The accuracy, detection rates, and false alarm rates may all have space for improvement, according to the data currently available. Earlier techniques like SVM and Naive Bayes can be replaced by other ones. The dataset can be improved by using particular procedures, according to the paper. to improve the input quality for the suggested system.

V. RESULT

The KDD dataset was used in the experiment conducted to test the proposed methodology, and the outcomes were pleasing. Our analysis is conducted using the following configurations: Hardware includes an Intel motherboard, an Intel core i3, 4 GB of RAM, and a 140 GB SSD drive. In terms of software, Windows 10 64-bit and Python 3.8. packages for Python like NumPy, pandas, and Kera Library KDD dataset: a set of data. For a normal class with recall, PCA and Random Forest both deliver good results.precession 0.99 0.099 and for attacker class 0.98 0.095. It also gave a pie chart analysis of test data for the type of attack.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 5 May 2023, pp: 856-859 www.ijaem.net ISSN: 2395-5252



VI. CONCLUSION:

An effective and efficient network intrusion detection system has been suggested in this paper. In this paper, we provide an intrusion detection approach that makes use of the Random Forest algorithm and Principal Component Analysis (PCA). In this case, the random forest will help with categorization while the PCA will help organise the dataset by reducing its dimensionality. As a result, it can be demonstrated that this intrusion detection system is more accurate and computationally quick.

REFERENCES:

[1]. JafarAbo Nada; Mohammad Rasmi Al-Mosa, 2018 International Arab Conference on Information Technology (ACIT), A Proposed Wireless Intrusion Detection Prevention and Attack System

[2]. Kinam Park; Youngrok Song; Yun-Gyung Cheong, 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigData Service), Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm

[3]. S. Bernard, L. Heutte and S. Adam "On the Selection of Decision Trees in Random Forests" Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009, 978-1-4244-3553-1/09/$25.00 ©2009 IEEE

[4]. A. Tesfahun, D. Lalitha Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction" 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 978-0-4799-2235-2/13 $26.00 © 2013 IEEE

[5]. Le, T.-T.-H., Kang, H., & Kim, H. (2019). The Impact of PCA-Scale Improving GRU Performance for Intrusion Detection. 2019 International Conference on Platform Technology and Service (PlatCon). Doi:10.1109/platcon.2019.8668960

[6]. Anish Halimaa A, Dr K.Sundarakantham: Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) 978-1-5386-9439-8/19/$31.00 ©2019 IEEE "MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM."

[7]. Mengmeng Ge, Xiping Fu, Naeem Syed, Zubair Baig, Gideon Teo, Antonio Robles-Kelly (2019). Deep Learning-Based Intrusion Detection for IoT Networks, 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 256-265, Japan.

[8]. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, "An Investigation on Intrusion Detection System Using Machine Learning" 978-1-5386-9276-9/18/$31.00 c2018IEEE.

[9]. Rohit Kumar Singh Gautam, Er. Amit Doegar; 2018 8th International Conference on Cloud Computing, Data

[10]. Science & Engineering (Confluence) "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms."

[11]. Kazi Abu Taher, Billal Mohammed Yasin Jisan, Md. Mahbubur Rahma, 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)"Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection."