iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42763

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Intrusion Detection System Using Principal Component Analysis with Random Forest Approach

J. Sanjay Rahul[1], J. Sai Keerthana[2], G. Tejaswini[3], R. N. S. Kalpana[4]
[1, 2, 3, 4, 5]*Electronics and Communication Engineering Dept, TKR College of Engineering and Technology, Hyderabad, Telangana, India*

Abstract: ***Through the evolution in radio communication, there are many security threats over the internet. The intrusion detection system helps to find the outbreaks on the system and the intruders are detected. Previously various machine learning techniques are applied on the IDS and tried to progress the results on the detection of trespassers and to increase the precision of the IDS. This paper has expected for an approach to develop IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to shape the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more resourcefully in terms of accuracy as compared to other methods like SVM, Naïve Bayes, and Decision Tree. The results obtained by projected method are having the values for performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and therefore the Error rate (%) is 0.21%.***

Keywords: ***Intrusion detection, Principal component Analysis, and Random Forest.***

I. INTRODUCTION

These days, the participation of the internet in regular life has been increased enormously. The internet has made a key role in everyone's life. The use of internet has become very crucial for everyone. So with the growth in the use of the internet for personal activities, it is also necessary to keep secure the system from malicious activities. Different outbreaks are seen on the system or the network. The attacks like a black hole, grey hole, warm hole etc are seen on the network system. These attacks are to snip the information from the system or to corrupt the data present over any system, to make misuse of the data, the intruders attack the system in various ways, some of the attacks are Dos, probe, snort, r21 etc. so to prevent the system from such attacks, the intrusion detection system was introduced. IDS keep track of attacks on the system and to prevent the system from these attacks. So to notice such attacks, the various methods have done earlier by using various techniques. Here an IDS that makes use of the principal component analysis is used along with the random forest technique. Both the methods work for a special purpose, where the PCA gives the roughness in the data, and the random forest helps the classification between the nodes for attacks.

II. METHODOLOGY

In this project we explore through Random forest, Intrusion detection, PCA models and test its performance.

A. Intrusion Detection System

Intrusion detection can be defined as the ability to monitor and react to computer misuse of the system without any permission and indulging the information present inside the system. This intrusion in any system can damage the hardware of the system. The intruder has become an important term to prevent the system from. The intruder inside any system can be controlled or keeping track of this intruder can be done with the help of the IDS. The numerous types of intrusion detection systems are used previous, but in the end, the accuracy concerns are seen in every method used. The two terms, such as detection rate and false alarm rate, are examined for the evaluation of the accuracy of the system. These two terms should be there in the system. The IDS is of two types in nature, for which it works are

1) *Network Intrusion Detection System:* In this system, the network traffic is studied, and the intrusion over this is estimated.
2) *Host-based Intrusion Detection System:* In this system keeps track of the system files that opened over the network. There is also subdivision of IDS types. The most common alternatives are based on signature detection and anomaly detection.
3) *Signature-Based:* In this, the system found some specific designs which are used by malware. These detected designs are called signatures. This is good in detecting known attacks, but when it comes to new attacks, it will be unsuccessful in such signature detection,

4) Anomaly-Base: This is specially developed for the detection of unidentified attacks.

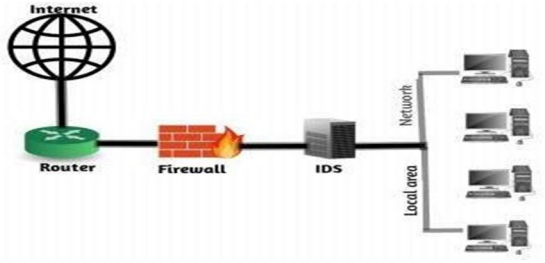


Figure: Intrusion detection system

B. PCA

Principal component analysis or PCA is dimensionality- reduction method which is used to reduce the dimensionality of the large dataset. The Principal component analysis is one of the well-organized 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 variables called principal component. This method takes all the input as the dataset, which is having a high number of attributes so the dimension of the dataset is very high. This technique reduces the size of the dataset by taking the datapoints on the same axis. This data points are shifted on a single axis, and the principal components are passed out. The PCA can be performed using the following steps:

1) Take the dataset with dimensions d.
2) Analyze the mean vector for each dimension d.
3) Compute covariance matrix for the entire dataset.
4) Compute eigen vector (e1, e2, …ed), and eigen vector(v1,v2…vd).
5) Perform sorting of eigen value in decreasing order and select n eigenvector with highest eigenvalues to get a matrix of d*n=M.
6) By this M forms a new sample space.
7) The outputs of sample spaces are the principal components.

C. Random Forest

Random forest is one of the most powerful device that is used in machine learning for classification problems. The random forest arises as the class of the supervised classification algorithm. This algorithm is passed in two different stages the first one deals with the construction of the forest of the given dataset, and the other one contracts with the prediction from the classifier that obtained in the very first stage.

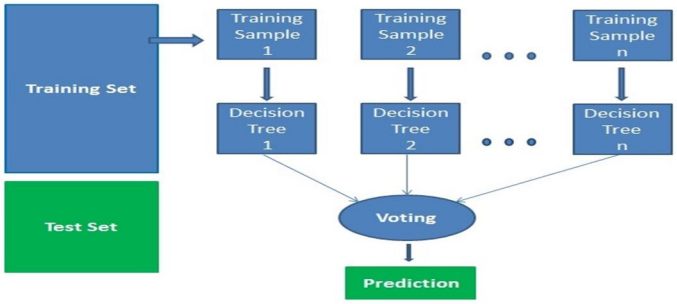


Figure: Random Forest

III. PROBLEM DOMAIN

The system which work done on the internet suffer from various malicious activities. The main problem seen in this field is the intruders in the system for violating the information. This intruder is noticed by an intrusion detection system. this system also needs to be exact and effective in the detection of the intruders. Numerous, machine learning algorithms were used for intrusion detection. Couple of those are: SVM, Naive Bayes, etc. But the consequences state that there may be some development to be done on terms of accuracy and the detection rates and the false alarm rate. Some other methods can substituted by earlier applied techniques such as SVM and Naïve Bayes. Also, the study states that the dataset can be upgraded by using some methods over it. To progress the quality of the input to the proposed system.

IV. PROPOSED SOLUTION

The intrusion detection system works for the development of the system, which is affected by the trespassers. This system can find the intruders. The proposed system tries to eradicate the existing problems related to the previous work. The proposed system involves in the two methods that are principal component analysis, and the other one is the random forest, The principal component analysis is used for the reduction of the dimension of the dataset, by this method the dataset quality will be enhanced as the dataset may contain the correct attributes. After this, the random forest algorithm will be applied for the detection of the intruders, which provide both the detection rate and the false alarm rate in an improved manner compared to SVM.

A. Algorithm for Proposed Solution

The attribute compatibility changes the coordination degree of the original attribute for the split node standard.

1) Attribute Compatibility: Authorize modulus be | Pr | for the core decision set, secondary set is | Se | & attribute compatibility is well-defined as:

$$CO(X \rightarrow D) = \frac{|pr| - |Se|}{|x|} \qquad \text{----(1)}$$

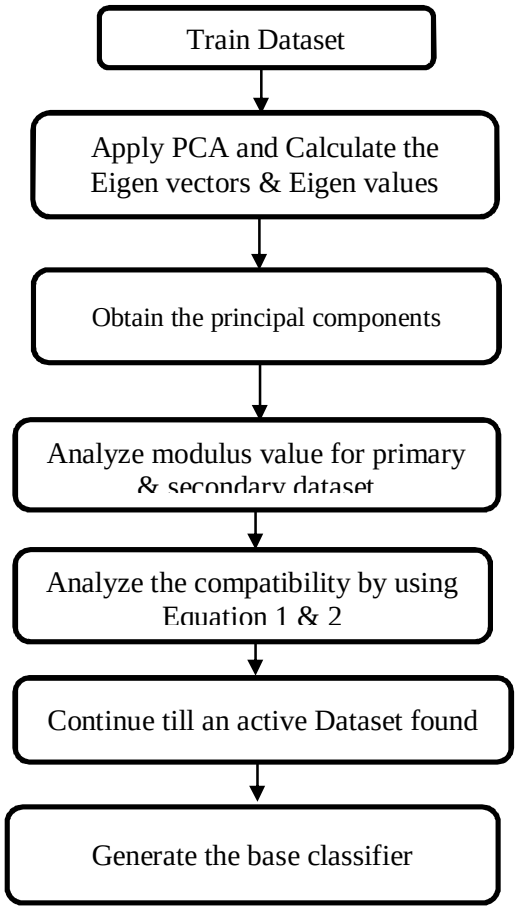
Here X, is subset for non-empty C.

2) *Strict Compatibility:* It is called after the influence of the secondary set over the mindset seen. A contradiction is seen between the main and the second set. The secondary set is round off by the expression:

$$CO(X \rightarrow O) = \frac{|Pr|}{|X|} \qquad \text{-----(2)}$$

Here X is subset for non-empty C. In this, varied compatibility of the second set is seen.

B. Flow Chart



V. RESULTS

The research carried out for the proposed method uses the KDD dataset, and the results obtained were satisfactory.

The following configurations are used for carrying our analysis:

1) Hardware requirements: 4GB RAM, 140Gb SSD Hard disk, Intel core i5

2) Software requirements: 64-bit windows10 and python

3) Python packages like NumPy, Pandas

4) Dataset: Kdd.

The application of PCA along with Random Forest operated well in comparison with existing techniques like SVM, Naive Bayes, and Decision tree.

The tabular represented below for the performance time(min), Accuracy rate (%), and Error rate (%) for different methods:

Technique	Performance time (min)	Accuracy rate (%)	Error rate (%)
SVM	4.56	84.33	2.66
Naïve Bayes	9.11	80.85	3.48
Decision tree	12.35	89.90	0.76
PCA with Random Forest	3.2	96.8	0.2

The table specified above gives an arithmetic representation of the obtained values from the research. The error rate obtained in our proposed approach is very low as of 0.2%. As well, the precision obtained is much higher than previous algorithms. Similarly, the time taken for the performance is less than additional algorithms.

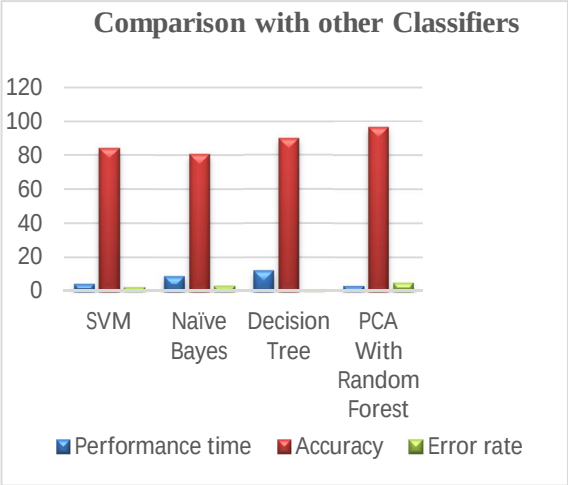


Figure: Comparison

VI. CONCLUSION

As per the involvement of the system over the internet increasing rapidly, the security concerns have also seen. The proposed approach deals with the recognition of the intruders over the internet efficiently. The proposed algorithm has performed well as related to the previously applied algorithms such as SVM, Naive Bayes, and Decision Tree. The detection rates and the false error rates can be enhanced at a great extent by the proposed approach. The dataset used in this is knowledge discovery dataset. The results obtained by our proposed techniques having the values for the performance time is 3.24min, Accuracy rate (%) is 96.77% and the Error rate (%) is 0.2%.

REFERENCES

[1] Hamzah nachann, dristi Poddar, Sambhaji sarode, dratik Kumhars "Intrusion detection system: A survey" by International Jounal of Engineering research is Technology, Issue 05, May-2021

[2] M.A Jabbar, Rajinikanth aluvalu "Intrusion Detection System using bayes Network and Feature subset solution", od Vardhaman college, Issue on 2017.

[3] Amol Borkar, Akshay Donode, Anjali Kumari "Survey on Intrusion detection system and Internal Intrusion detection and Protection system" of computer Engineering, Issue on 2017.

[4] Madhukar, Nantha kumar "An Intruder Detection system based on feature selection using Random Forest" by Internal Journal of engineering and Advanced Technology, Issued 02, December-2019

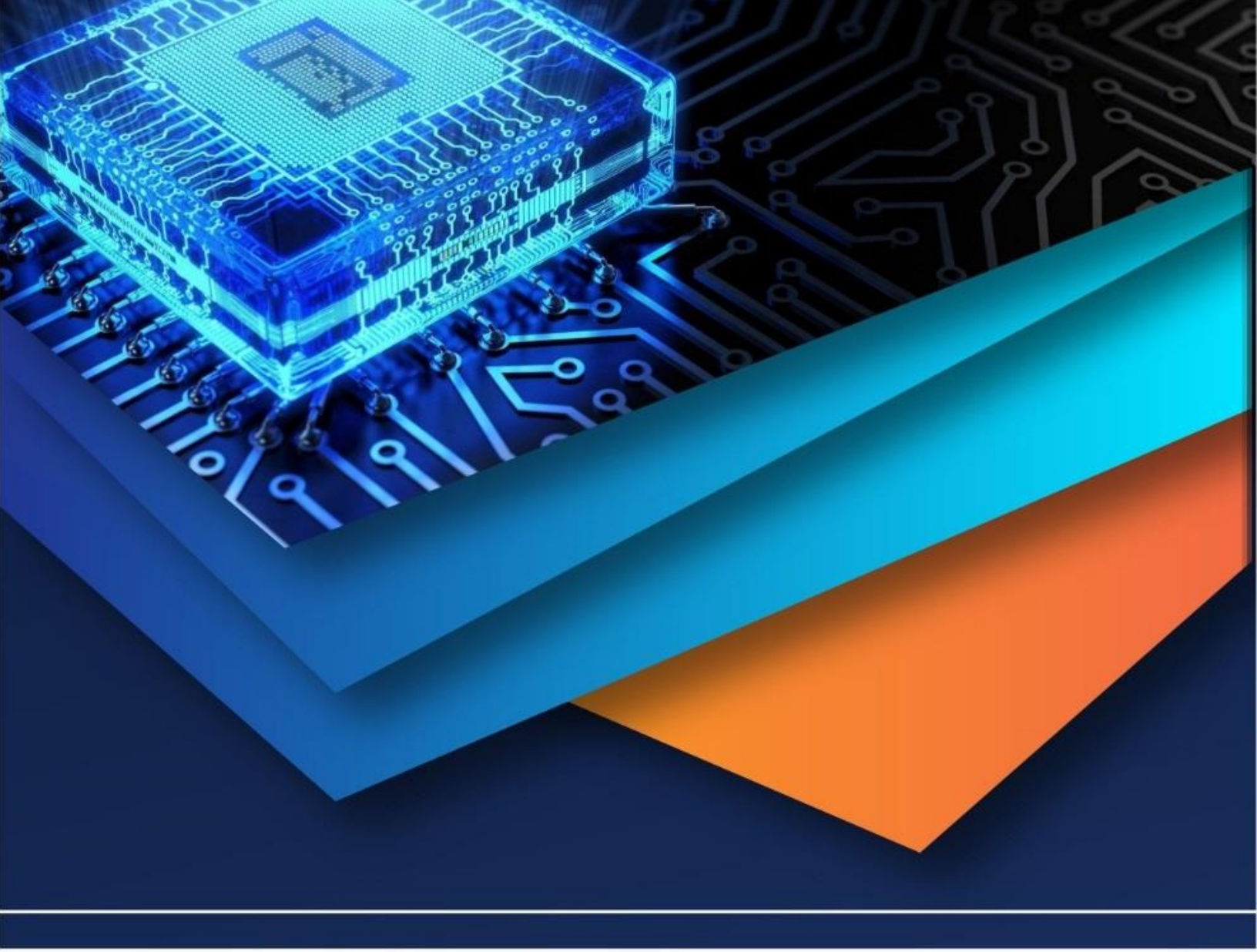
AUTHORS

R.N.S. KALPANA received her B. Tech degree in Electronics and Communication Engineering from JNTU in 2007 and she completed her M. Tech from ARORA HYDRABAD in 2010. Presently she is working as Assistant Professor in TKR college of Engineering and Technology, HYD.

J. SANJAY RAHUL pursuing his B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology, 2021-2022, HYD

J.SAI KEERTHANA, pursuing her B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology, 2021-2022, HYD.

G. TEJASWINI, pursuing her B. Tech final year in Electronics and Communication Engineering from TKR college of Engineering and Technology, 2021-2022, HYD.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY