International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Intrusion Detection System Using PCA with Random Forest Approach

SOMARAJU RAM PRASAD[1], POTHINA AMBIKA[2], K. MOUNIKA[3]

[1,2,3] *Final Year B.Tech, CSE, Sanketika Vidya Parishad Engineering College, Visakhapatnam, A.P, India Guided by G. Geetha vaishnavi, Associate Professor, SVPEC, Visakhapatnam, A.P, India*

Abstract - *With the evolution in wi-fi communication, there are numerous protection threats over the net. The intrusion detection system(IDS) enables the invention of the assaults on the gadget and therefore the intruders are detected. Previously numerous system learning (ML) strategies are dole out at the IDS and attempted to reinforce the results of the detection of intruders and to boom the accuracy of the IDS. This paper has proposed a way for IDS through the usage of the principle component analysis (PCA) and also the random forest algorithm. Where the PCA will assist to organise the dataset by lowering the dimensionality of the dataset and therefore the random forest will assist in type. Results acquired state that the proposed method works extra effectively in phrases of accuracy compared to different strategies like SVM, Naïve Bayes, and Decision Tree. the consequences acquired through the proposed technique are having the values for overall performance time (min) is 3.24 minutes, Accuracy rate (%) is 96.78 %, and therefore the Error rate (%) is 0.21 %.*

***Key Words*: — IDS, Knowledge Discovery Dataset, PCA, Random Forest, SVM, Naïve Bayes, and Decision Tree.**

1. INTRODUCTION

Nowadays, the involvement of the web in normal life has increased rapidly. the web has made an important place in everyone's life. the employment of the web has become very crucial for everybody. So with the rise within the use of the web for private activities, it's also necessary to stay secure the system from malicious activities. Different attacks are seen on the system or the network. The attacks sort of a region, grey hole, wormhole, etc. are seen on the network system. These attacks are to steal the knowledge from the system or to corrupt the information present over any system. to form misuse the info, the intruders attack the system in various ways, a number of the attacks are DoS, probe, snort, r2l, etc. So to forestall the system from such attacks, the intrusion detection system was introduced. IDS keeps track of attacks on the system and prevents the system from these attacks. So to detect such attacks, the assorted works are done earlier by using various techniques. Here an intrusion detection system that creates use of the principal component analysis is employed together with the random forest technique. Both the methods work for a special purpose, where the PCA gives the granularity within the data, and also the random forest helps the classification between the nodes for attacks.

1.1 intrusions Detection System:

Intrusion could be a term that deals with entering the system with none permission and spoiling the data present inside the system. This intrusion in any system may also harm the hardware of the system. The intrusion has become a major term to forestall the system from. This intrusion inside any system are often controlled or keeping track of this intrusion are often through with the assistance of the IDS. the varied styles of intrusion detection systems are used earlier, but within the end, the accuracy concerns are seen in every method used. the 2 terms, like detection rate and also the warning rate, are analyzed for the evaluation of the accuracy of the system. These two terms should be within the manner that the warning rate should be minimized and also the improvement within the detection rate should be there within the system. therefore the random forest together with the PCA is applied for the IDS. The IDS may be of two types in nature, that it works, that are:

- Network Intrusion Detection Systems (NIDS): during this system, the network traffic is analyzed, and also the intrusion over it's analyzed.
- Host-based Intrusion Detection Systems (HIDS): Here, the system keeps track of the system files that are accessed over the network. there's also a subset of IDS types. the foremost common variants are supported signature detection and anomaly detection.
- Signature-based: during this, the system found some specific patterns which are utilized by malware. These detected patterns are called signatures. this is often good in detecting known attacks, but when it involves new attacks, it fails in such signature detection.
- Anomaly-based: this is often specially developed for the detection of unknown attacks. this method uses ML to construct the model

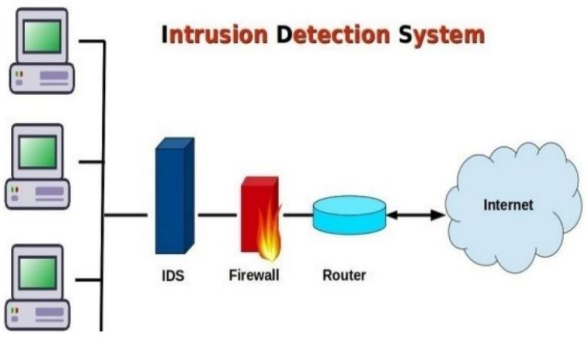

Figure 1. Intrusion Detection System [2]

2. RELATED WORK

Intrusion could be a period of time that provides to urge into the machine with statistics within the machine. This intrusion into any machine can also damage the hardware of the machine. It's grown to be a motivating period to save lots of you the machine. This intrusion inner any machine might be managed or perhaps retaining the song of this intrusion could also be achieved with the help of the IDS. the various kinds of intrusion structures are used earlier, however, withinside the tip, the accuracy worries are apparent in each approach used.

The phrases, inclusive of detection price and therefore the fake alarm price, are analyzed for the assessment of the accuracy of the machine. These phrases need to be withinside the way that the fake alarm price needs to be minimized and therefore the development withinside the detection price has got to be there withinside the machine. therefore the random woodland at the sting of the PCA is administrated.

Random Forest:

Random Forest is that the prevalent supervised technique. it's useful for mainly doing classification challenges and also regression challenges. RF is one amongst the classifiers which holds multiple decision trees in each subset of an assumed data set and computes the everyday value that enhances prediction accurateness for the dataset.

The random forest doesn't depend upon decision trees. Instead, it gets a prediction from every tree so forecasts the last result which is made upon polls of prevalence estimations. The more trees within the forest, the upper the accuracy and avoid overfitting problems. it's supported the ensemble technique concept, which mixes multiple classifiers to unravel a thorny problem and improves model performance.

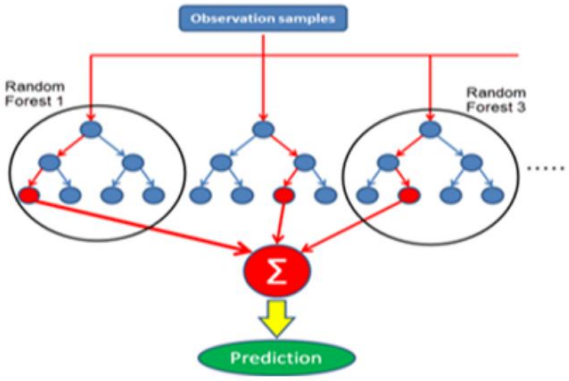

Figure 2. Random Forest Model.

Principal component analysis:

The principal component analysis is that the technique that's used, significantly for the reduction of the dimension of the given dataset. The principal component analysis is one amongst the foremost efficient and detailed methods for reducing the scale of information, and it provides the specified results [6]. This method reduces the aspects of the given dataset into a desired number of attributes called principal components. This method takes all the input because the dataset, which has a high number of attributes therefore the dimension of the dataset is extremely high. This method reduces the dimensions of the informationset by taking the data points on the identical axis. the info points are shifted on one axis, and therefore the principal components are dole out. The PCA will be performed using the subsequent steps:

1. Take the dataset with all dimensions d.
2. Calculate the mean vector for every dimension d.
3. Calculate the covariance matrix for the entire dataset.
4. Calculate the eigen vectors (e1, e2, e3 … .ed), and eigen values (v1,, v2, v3, ….vd).
5. Perform sorting of eigenvalue in decreasing order and choose n eigenvector with the best eigenvalues to induce a matrix of d*n= M.
6. By using this M form a replacement sample space.
7. The obtained sample spaces are the principal components

3. PROBLEM DOMAIN:

The systems which beat up the web suffer from various malicious activities. the foremost problem seen during this field is that the intrusion into the system for violating the

knowledge. This intrusion is caught by creating an intrusion detection system; this technique also has to be accurate and efficient within the detection of intruders. Various machine learning algorithms were used for intrusion detection; a number of them are SVM, Naïve Bayes, etc. But the outcomes state that there is also some advancements to be wiped out terms of accurateness and also the detection rates and therefore the warning rate. another approaches can replace previously applied procedures like SVM and Naïve Bayes. Also, the study states that the dataset is enhanced by using some approaches over it. to reinforce the standard of the input to the proposed system.

4. PROPOSED SOLUTION:

The intrusion detection machine works for the event of the machine, that's experiencing the intruders. This device can do the detection of intruders. The proposed machine attempts to get rid of the prevailing issues related to the preceding work. The proposed machine includes the 2 techniques which are important aspect evaluation, and therefore the opposite one is that the random woodland. The key aspect evaluation is employed for the discount of the dimensions of the dataset; with the help of creating use of this method, the dataset first-rate can be advanced because the dataset might also incorporate probably the simplest attributes. Next, the random woodland set of rules could also be applied for the detection of the intruders, which give each the detection price and also the fake alarm price in a sophisticated way compared to SVM.

4.1. Algorithm for the proposed solution:

The attribute compatibility substitute the coordination degree of the actual attribute for the split node standard.

1. Attribute compatibility Let the modulus be | Pr | for the main decision set, secondary set be | Se |, and attribute compatibility is defined as:

$$CO(X \rightarrow D) = \frac{|P_r| - |S_e|}{|X|} \quad (1)$$

Here, X, is that the subset for non-empty C. Strict compatibility is termed when the impact of the secondary set over the mindsets is seen. A contradiction is seen between the most and also the second set. The secondary set is rounded off by the expression

$$CO(X \rightarrow D) = \frac{|P_r|}{|X|} \quad (2)$$

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

Algorithm for the bottom Classifier Improvement:

Step 1: initialisation of information set active attribute by marking all condition attributes.

Step 2: calculate the modulus for each condition attribute in both primary and secondary set.

Step 3: By using equation (1) compatibility calculation of all conditional attribute is completed during this step. Use equation (2) if more characteristic with similar compatibility is seen.

Step 4: To separate the sample, select the foremost extensive compatibility for splitting because the split node and delete the active tag.

Step 5: persist selecting the active attribute for splitting till the active quality is reached up to leaf node.

Step 6: finally, the bottom classifier is generated.

4.2. Flowchart for the Proposed Algorithm:

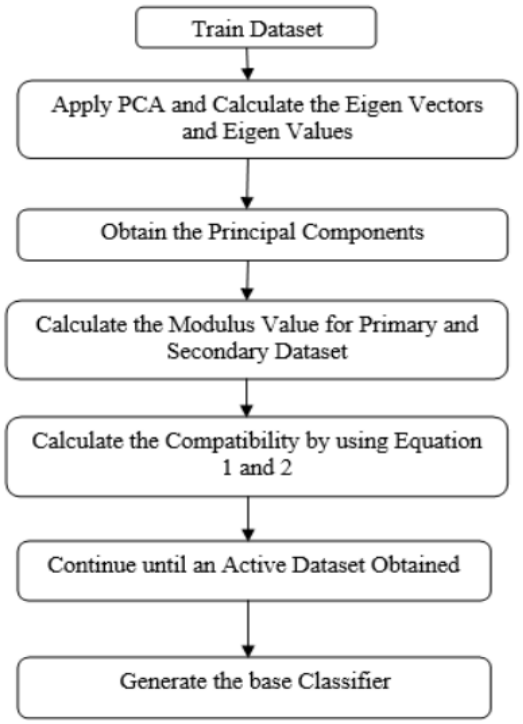

Figure 3. Flowchart for the Proposed Approach

5. RESULTS

The experiment was administrated for the distinctive machine learning models and KDD dataset, and therefore the results obtained, Intrusion Intrusion Detection Systems (IDS) are among the ways against these attacks.

Furthermore, modern technologies of upcoming generation networks such as an example Wireless networks normally stated as Wi-Fi have appeared, which imply a notable comprehension of the key difficulties and constraints that house the layout similarly because the setup of an IDS for such methods. IDS often needs to boost its performance of its in conditions of raising the precision and lessening false alarms. In machine learning grounded IDS, integrating effective feature selection in addition as attribute dimensionality minimization with intrusion detection has demonstrated to be a booming strategy since it can assist in choosing probably the foremost informative features and minimize the function dimensionality from the complete set of characteristics. Here, X, is that the subset for non-empty C. Strict compatibility is termed when the impact of the secondary set over the mindsets is seen. A contradiction is seen between the most and also the second set. The secondary set is rounded off by the expression

Table 2. Result Comparison with other Classifiers

Method	Performance time (min)	Accuracy rate (%)	Error rate (%)
SVM	4.57	84.34	2.67
Naïve Bayes	9.12	80.85	3.49
Decision Tree	12.36	89.91	0.78
PCA with Random Forest	3.42	96.78	0.21

The table given above gives a numerical presentation of the acquired values from the experiment. The error rate found in our proposed approach is incredibly low at .21%. As well, the accuracy acquired is far more elevated than in earlier algorithms. Also, the time taken for the performance is a smaller amount than other algorithms.

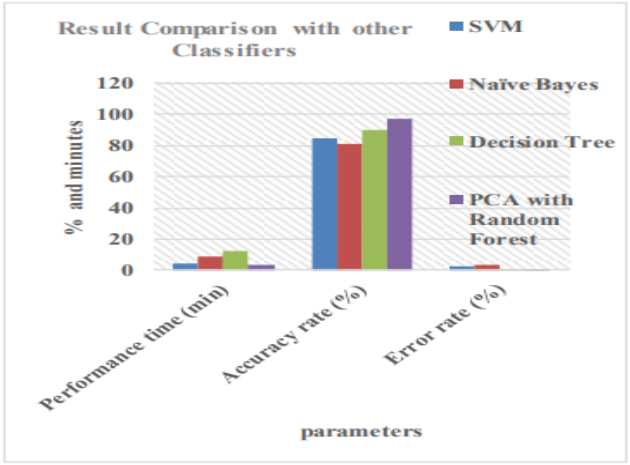

Figure 4. Result Comparison with other Classifiers

6. CONCLUSION

As the participation of the devices over the online grows quickly, questions of safety are found. The suggested answer pertains to the detection of intruders over the online effectively. The advised set of rules has been concluded thoroughly as whilst compared to the earlier executed algorithms like SVM, Naïve Bayes, and Decision Tree. The detection prices in complement to the factitious blunders prices may be substantially improved at an exquisite quantity with the help of using the advised answer. The results obtained with the help of using maintaining the values for a Performance period(min)is 3.24mins, the Accuracy rate(%) is 96.78 %, and therefore the Error fee (%) is 0.21 %.

REFERENCES:

1. Jafar Abo 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 (BigDataService), Classification of Attack Types for Intrusion Detection Systems Using a Machine Learning Algorithm 3

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 Science & Engineering (Confluence) “ An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms.”

10. 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.”

11. L. Haripriya, M.A. Jabbar, 2018 Second International Conference on Electronics, Communication and Aerospace

    Technology (ICECA)” Role of Machine Learning in Intrusion Detection System: Review”

12. Nimmy Krishnan, A. Salim, 2018 International CET Conference on Control, Communication, and Computing (IC4) “ Machine Learning-Based Intrusion Detection for Virtualized Infrastructures”

13. Mohammed Ishaque, Ladislav Hudec, 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)“Feature extraction using Deep Learning for Intrusion Detection System.”

14. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)“A Review of Machine Learning Methodologies for Network Intrusion Detection.”

15. Iftikhar Ahmad , Mohammad Basheri, Muhammad Javed Iqbal, Aneel Rahim, IEEE Access (Volume: 6) Page(s): 33789 – 33795 “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection.”

16. 98 B. Riyaz, S. Ganapathy, 2018 International Conference on Recent Trends in Advanced Computing (ICRTAC)” An Intelligent Fuzzy Rule-based Feature Selection for Effective Intrusion Detection.”

BIOGRAPHIES

G. GEETHA VAISHNAVI

Currently working as associate professor from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering

SOMARAJU RAM PRASAD

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

POTHINA AMBIKA

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

K.MOUNIKA

Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College