iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44326

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 VI June 2022- Available at www.ijraset.com

Network Intrusion Detection System Using Machine Learning with Data Preprocessing and Feature Extraction

Manvith Pallepati[1], Soujenya Voggu[2], Rithesh Masula[3], Manisai Konjarla[4]
[1, 2, 3, 4]*Sreenidhi Institute of Science and Technology*

Abstract: ***Unauthorized access to a computer network can be discovered by scanning the network traffic for evidence of malicious activity, which is what Network Intrusion Detection (NID) does. However, in this study, we will concentrate on the technology, development, and strategic importance that make up the large field of Network Intrusion Detection (NID). Many new strategies have been created in the last few years to help computer security specialists in protecting a single host or an entire network against unauthorized access, theft, and denial-of-service assaults, which are the primary causes of computer crime. Intrusion Detection is critical for both the military and commercial sectors since it is the most significant study area for the future networks' Information Security. In this paper, a model is being proposed, where the data is preprocessed before training with the algorithms. A study done by comparing with other models shows that, the current model built with Random Forest can outperform other existing models built with ANN when the data is preprocessed. After building model after data pre-processing and feature extraction, we are able to achieve 98.71% accuracy on NSL-KDD dataset.***

Keywords: Network intrusion detection (NID), random forest, multi-layer perceptron are some of the terms used in this article.

I. INTRODUCTION

Network security has become a major problem in today's world, as the unlawful actions in the networking world continue to rise at an alarming rate, making network security neither hopeless nor addressed. There was a time when only firewalls were in place to protect networks from cyberattacks; Morris Worm launched the first internet-wide attack and penetration on November 2, 1988. Since then, new technologies have been created to protect networks from cyberattacks, as well. Vendors used to ship the user name and password along with the equipment in order to avoid the danger of a hacker gaining access. In the mid-1980s, Denning's work on Intrusion Detection (ID) was published for the first time in print. Because intrusive activity differs from regular behavior, Denning believes that the primary role of the IDS is to develop acceptable models of normal behavior that may be used to identify intrusive activity. In computer networking, intrusion detection is the most recent and essential study subject. A number of prototypes have been created that use a variety of methodologies for commercial and military uses alike, with ID being one of the most popular topics of discussion.

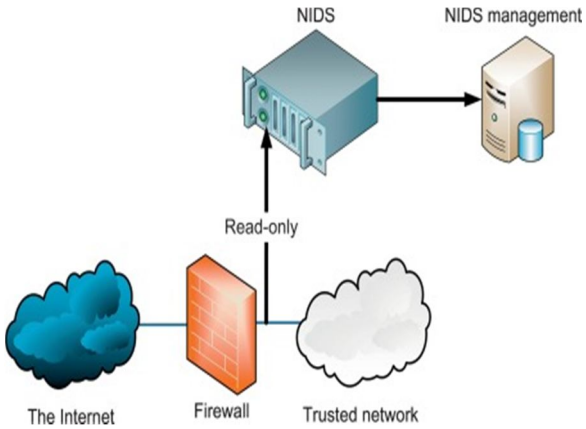


Fig.3: Network intrusion detection system

©IJRASET: All Rights are Reserved

ID relies on a mix of brute force and incursion tactics. An IDS (Intrusion Detection System) is now an essential aspect of network security since it protects the network from both successful and unsuccessful intrusions. Detecting and reporting any unexpected system behavior is the primary goal of the intrusion detection system (IDS). IDS can provide a real-time response to all intrusion events after a thorough investigation of the behavior and signature of the intrusion detection. As well as the above, IDS is capable of

• Verification of system errors; monitoring of system and user activity.

System and data file integrity is being assessed.

• Keep a running tally of any unusual conduct.

• A model for mapping known assaults and alerts to recognize them.

Some IDS respond to an illegal intrusion in the system by terminating the network connection as a response to the intrusion, rather than just preventing it from happening in the first place [2].

For IDS, the following are the primary criteria for measuring performance.

• TRUE POSITIVE: an actual attack (IDS gives alarm).

An attack has not taken place (IDS gives alarm).

FALSE NEGATIVE: a smear campaign (IDS gives no alarm)

NEGATIVE VERBAL: there was no attack (IDS gives no alarm)

A. NSL-KDD Dataset

The NSL-KDD dataset consists of attacks for 4 different categories with 43 features for each attack. Where 41 features refers directly to the network traffic input. This dataset is used as standard dataset for network intrusions. The dataset contains 39 different subclass types of the attacks.

Classes:	DoS	Probe	U2R	R2L
Sub-Classes:	• apache2 • back • land • neptune • mailbomb • pod • processtable • smurf • teardrop • udpstorm • worm	• ipsweep • mscan • nmap • portsweep • saint • satan	• buffer_overflow • loadmodule • perl • ps • rootkit • sqlattack • xterm	• ftp_write • guess_passwd • httptunnel • imap • multihop • named • phf • sendmail • Snmpgetattack • spy • snmpguess • warezclient • warezmaster • xlock • xsnoop
Total:	11	6	7	15

Fig.1: NSL-KDD attacks with subclasses

Dataset	Number of Records:					
	Total	Normal	DoS	Probe	U2R	R2L
KDDTrain+20%	25192	13449 (53%)	9234 (37%)	2289 (9.16%)	11 (0.04%)	209 (0.8%)
KDDTrain+	125973	67343 (53%)	45927 (37%)	11656 (9.11%)	52 (0.04%)	995 (0.85%)
KDDTest+	22544	9711 (43%)	7458 (33%)	2421 (11%)	200 (0.9%)	2654 (12.1%)

Fig.2: NSL-KDD Dataset

II. LITERATURE REVIEW

A. Technology and development of intrusion detection systems:

Currently, attacks on network infrastructure pose the most serious dangers to the safety of networks and their associated data. Because standard firewall techniques cannot provide comprehensive protection against intrusion, intrusion detection (ID) has become an increasingly important part of defense-in-depth strategies to keep networks free of illegal activity. Network security's ID field is one that's constantly being worked on. This report presents the results of a survey on biometric identification technologies. It has a hand in a number of critical facets of the identification process. Data collection and intrusion detection strategies are examined in depth. Data mining-based and data fusion-based approaches to ID systems are also explored in this article. The current state of ID technology is confronted with formidable obstacles, and both current difficulties and possible future prospects are discussed.

B. Intrusion Detection systems: Technology and Development

A computer network's intrusion detection system is one of its primary lines of defense. There has been a lack of concentration and direction in this study in recent years [3]. Multiagent intrusion detection systems have made significant strides in recent years, and we examine those advances in this study. We do a review of the existing Intrusion Detection System types, strategies, and architectures in the literature. Finally, we provide an overview of the current state of research.

C. Use Of Data Mining And String Metrics For String Anomaly Detection

Cyber-attacks on computers and networks are becoming more frequent and more severe. In the modern world, intrusion detection is a vital technology and an active research topic. This research presents an adaptive method to anomaly-based intrusion detection systems using data mining techniques and string metrics. Based on the findings of the simulation studies, the proposed method is proven to generate trustworthy results while simultaneously monitoring and alerting the protected system.

D. The Role Of Intrusion Detection Systems In Protecting Yourself

Systems for detecting intrusions into computer networks and systems are an important part of the defenses in place to keep them safe. Using IDSs as part of a company's overall defensive posture is discussed in this article, as are best practices for IDS deployment, operation, and maintenance.

E. A Data Mining-Based Study On Intrusion Detection

It's important for every company to keep their data safe from both internal and external threats. The first line of protection in this project is firewall, encryption, and authentication. The second line of defense is Intrusion Detection. Anomaly-based or misuse-based approaches can be used by IDS. There was a time when traditional IDS relied on a misuse-based approach. An issue with misuse-based approaches is that they are not able to identify new types of assaults. As a result, anomaly-based intrusion detection was employed, which can pick up on attempts that aren't immediately obvious. Because of the advantages data mining brings, anomaly-based intrusion detection makes heavy use of it. In this research, we use data mining to investigate the aspects of intrusion detection.

III. IMPLEMENTATION

It has been more critical in recent years than ever before to have an effective intrusion detection system. Current intrusion detection systems have a number of issues, such as slow network speed, lack of support for IPv6 addressing scheme, high false positive and false negative alarm rates, absence of intrusion detection benchmarking, and lack of accuracy in real-time detection.

Inconsistencies in the Current System:

1) Cost of the System

2) Not entirely prohibited

A. Plans for a New System

A network detection mechanism is proposed in this article. An IDS (Intrusion Detection System) is now an essential aspect of network security since it protects the network from both successful and unsuccessful intrusions. The IDS's primary function is to detect and report on any and all aberrant system behavior, including non-attacks. An in-depth examination of intrusion detection behaviors and signatures will allow IDS to provide a real-time response to any and all intrusion occurrences.

The following are some of the proposed system's advantages:

- Improved foresight.
- Lower cost

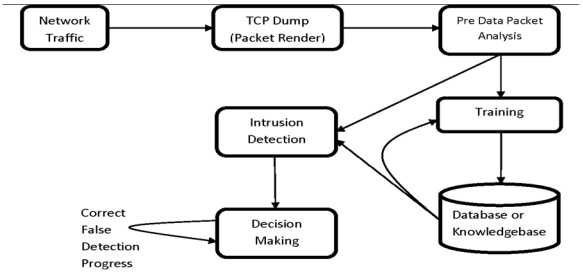


Fig.2: System architecture

B. Modules

1) Pre-processing of the data It is possible to normalize data using the standard scalar Using One-Hot-Encoding to convert Categorical data to Numeric data
2) Extraction of Characteristics (Features)
3) Finding out which of the following attributes has a Pearson-correlation of higher than 0.5 with the encoded attack label attribute:
4) One-hot encoded and original attack labels can now be combined.
5) model's training algorithms (Training data: 75 % , Testing data: 25 %)
6) Then Training with MLP (Multilayer Perceptron Algorithm)
7) Same data set Training with Random Forest Algorithm

An IDS based on signatures will scan the network for malicious packets and compare them to a database of known signatures. By scanning each intrusion event for a known attack or signature, they work similarly to a virus scanner [5]. While a signature-based IDS is extremely effective at detecting known attacks, it does, like anti-virus software, require regular updates to keep up with changes in hacker techniques. This strategy makes it simple to spot the most prevalent and well-known types of attacks. However, one of the primary issues is how to construct a signature that encompasses all the modifications and variations of the assault. If the attack pattern is fully unknown, this strategy will not work. Changing the attack signature is not an option. An open source signature-based IDS known as SNORT has been used as a benchmark by numerous academics to compare their own IDS to. As the system is being built on Random Forest, which is supervised learning algorithm, it doesn't need heavy GPU's to run. Hence, the cost of the model is being reduced.

IV. ALGORITHM

A. MLP

A feedforward artificial neural network, a multilayer perceptron (MLP) produces a collection of outputs from a set of inputs. Directed graphs are used to connect the input and output layers of an MLP. In order to train the network, MLP employs back propagation. Multiple layers of neurons are placed together in a Multilayer Perceptron to form input and output layers and one or more hidden levels. Neurons can employ any activation function in a Multilayer Perceptron, unlike neurons in a Perceptron, which must have a threshold-imposing activation function like ReLU or sigmoid.

The layers are hidden in multi-layer perceptron (other than a single input and single output layers). A multi-layer perceptron, in contrast to a single-layer perceptron, may learn non-linear functions as well as linear ones.

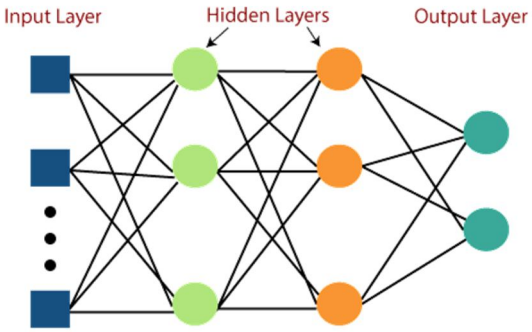


Fig.3: MLP model

The perceptron is only capable of solving simple issues. It has the ability to insert a hyperplane in pattern space and move it till the mistake is minimized. Only if the problem can be linearly divided does this method work.

B. Random Forest

An algorithm known as Random Forest is a Supervised Machine Learning Algorithm that is commonly utilized in classification and regression applications. It creates decision trees on a variety of samples and uses their majority vote to classify and average the data. For classification and regression, random forests and random decision forests are both ensemble learning methods that work by training a large number of decision trees at the same time. The random forest's output is the categorization chosen by the majority of trees when performing classification tasks.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Individual tree predictions are averaged together for regression tasks. Overfitting of decision trees by random decision forests is corrected. To some extent they beat decision trees, but random forests are less accurate than gradient-boosted trees. Data features, on the other hand, have the potential to influence their performance.

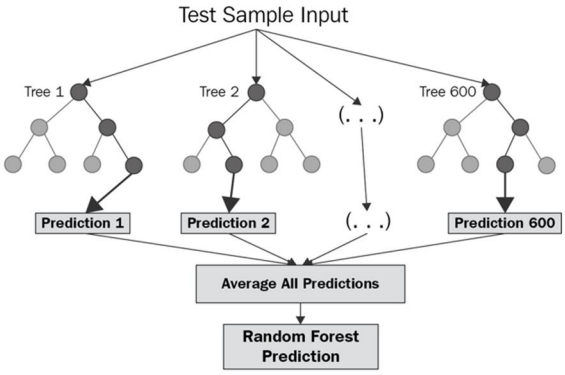


Fig.4: Random forest model

In addition to regression and classification tasks, it can also be used. The forecasts produced by a random forest are accurate and understandable. It is able to process massive amounts of data quickly. The random forest method outperforms the decision tree algorithm when it comes to forecasting outcomes.

V. RESEARCH AND DEVELOPMENT OF INTRUSION & DETECTION

Among the many methods, some of which are outlined in the following paragraphs:

A. IDS On the Basis of Data Mining

This technique, known as data mining, is based on randomly extracting hidden predictive information from huge, incomplete, noisy and confusing datasets. In a data-mining strategy, features are extracted from a big quantity of data automatically [8]. For mining audit data, the following techniques are employed to uncover previously unseen relationships among a plethora of variables:
When it comes to categorization, It's important to conduct link analysis. Analysis of the sequence of events
There is only one drawback to using audit trails: the amount of data required for analysis is prohibitively expensive because of the volume. On the other hand, despite its many advantages, the approach is still in its infancy and requires a great deal of research.

B. Integrated Data Security (IDS)

We can utilize the multiple sensors fusion methodology for intrusion detection because it is a relatively new way for integrating data from numerous sources and sensors [9]. With the use of historical data, data fusion was able to determine the current state of the network. Fusion-based techniques face an important challenge: how to merge multiple data sources from a network. In addition to the Bayesian approach, data fusion techniques are expected to be one of the most active study areas in the near future.
An IDS is only as good as the data it collects. The following data is collected at various levels:
• Data from the network
• The operating system's request in the context of any software program
• Keyboard input
In the operating system's command line
Host-based data refers to systems that collect operating system data that is vulnerable to attack. Observing network traffic for signs of intrusion is an alternative to host-based sensing, which focuses on the system or systems being watched. Using this technology, a single security sensor may monitor a large number of hosts and hunt for attacks that affect numerous hosts. As a result, assaults like those sent from a system console or a dial-up modem physically connected to the host are invisible.
The following services should be provided via secure and protected network systems.
• Data security is a primary concern.
Consistency in information transmission and storage
• Protection against DDoS assaults

In order to protect the integrity of the data, data confidentiality ensures that no unauthorized individual can access it. An important part of data and communications integrity is ensuring that information is transmitted accurately, without causing damage to individual pieces of data, and that it can be trusted. All of this service must ensure that the hardware and software of the system work perfectly, and that the data is protected from being tampered with. A common type of website hacker attack, denial-of-service is used to take down an entire website at once. When a (remote) entity cannot be accessed, this happens. Despite the fact that these attacks are not fully prevented, we can at least lessen their likelihood. To keep a computer or network safe, most people use security settings to create a protective shield around it. To get access to the system, intruders must identify and validate their identity via a series of security checks.

VI. EXPERIMENTAL RESULTS

```
y_pred = rf_mod.predict(X_test)# predicting target attribute on testing dataset
ac = accuracy_score(y_test, y_pred)*100 # calculating accuracy of predicted data
print("Random Forest -Accuracy is ", ac)

[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.

Random Forest -Accuracy is  98.7140407696704

[Parallel(n_jobs=1)]: Done 200 out of 200 | elapsed:    0.7s finished
```

Fig.5: Random forest accuracy

Random Forest accuracy achieved is 98.7140%

```
# defining loss function, optimizer, metrics and then compiling model
mlp.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# predicting target attribute on testing dataset
test_results = mlp.evaluate(X_test, y_test, verbose=1)
print(f'Test results - Loss: {test_results[0]} - Accuracy: {test_results[1]*100}')

985/985 [==============================] - 1s 1ms/step - loss: 0.0661 - accuracy: 0.9775
Test results - Loss: 0.06609897315502167 - Accuracy: 97.74878025054932
```

Fig.6: MLP accuracy

MLP accuracy achieved is 97.7487%

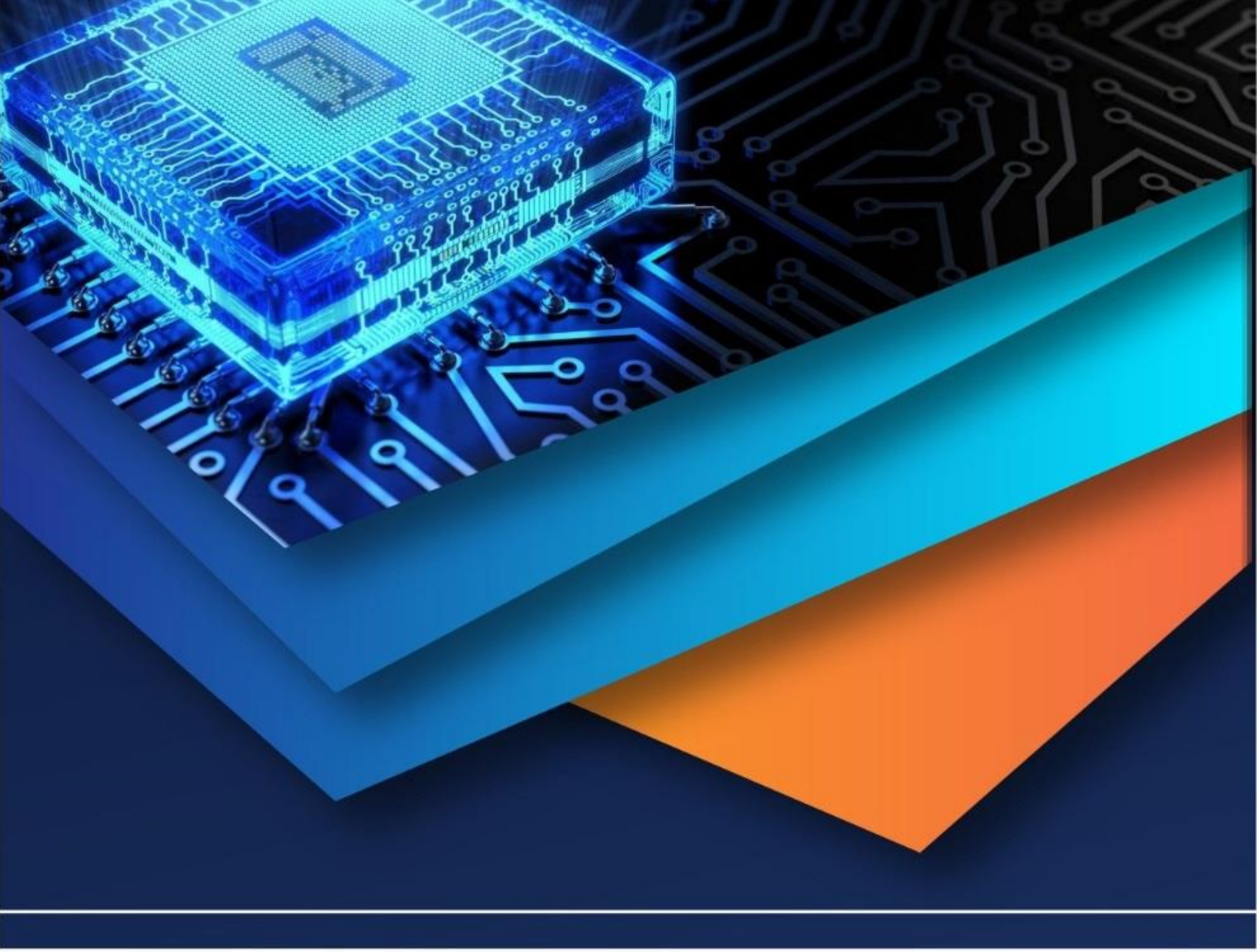
VII. CONCLUSION

Here, an introduction to Intrusion Detection technology is given as well as a look at the current technology, its challenges, and the future of Intrusion Detection technology are discussed in detail. Currently, intrusion detection is a relatively new technology, which means that there are a huge number of open research, engineering, and scientific prospects. Intrusion Detection Systems are widely utilized around the world, and more money is being invested in research and development of new systems. Intrusion Detection technologies will continue to improve since the threat of a larger attack will never go away.

Hence, we can conclude that the model built with Random Forest outperformed the model built using ANN techniques. Also the cost of the model is reduced by huge as the Random Forest doesn't require GPU's to run.

REFERENCES

[1] D. E. Denning, "An intrusion-detection model," Software Engineering, IEEE Transactions on, no. 2, pp. 222–232, 1987.
[2] Y. Bai and H. Kobayashi, "Intrusion detection systems: technology and development," in Advanced Information Networking and Applications, 2003. AINA 2003. 17th International Conference on. IEEE, 2003, pp. 710–715.
[3] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," Network, IEEE, vol. 8, no. 3, pp. 26–41, 1994.
[4] M. F. Marhusin, D. Cornforth, and H. Larkin, "An overview of recent advances in intrusion detection," in Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on. IEEE, 2008, pp. 432–437.
[5] E. Nikolova and V. Jecheva, "Anomaly based intrusion detection using data mining and string metrics," in Communications and Mobile Computing, 2009. CMC'09. WRI International Conference on, vol. 3. IEEE, 2009, pp. 440–444.
[6] J. McHugh, A. Christie, and J. Allen, "Defending yourself: The role of intrusion detection systems," Software, IEEE, vol. 17, no. 5, pp. 42–51, 2000.
[7] D. Dasgupta, "Advances in artificial immune systems," Computational Intelligence Magazine, IEEE, vol. 1, no. 4, pp. 40–49, 2006.
[8] S. Naiping and Z. Genyuan, "A study on intrusion detection based on data mining," in Information Science and Management Engineering (ISME), 2010 International Conference of, vol. 1. IEEE, 2010, pp. 135–138.
[9] M. Shankar, N. Rao, and S. Batsell, "Fusing intrusion data for detection and containment," in Military Communications Conference, 2003. MILCOM'03. 2003 IEEE, vol. 2. IEEE, 2003, pp. 741–746.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)