IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 3, March 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

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

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

Intrusion Detection of Imbalanced Network based on Machine Learning Algorithm: A Review

Amrita Singh[1], Dr. Vijay Bhandari[2], Dr. Ritu Srivastava[3]

M. Tech. Scholar, Dept. of Computer Science and Engineering, Sagar Institute of Research Technology & Science, Bhopal, India[1]

Associate Professor, Dept. of Computer Science and Engineering, Sagar Institute of Research Technology & Science, Bhopal, India [2]

Head of Dept., Dept. of Computer Science and Engineering, Sagar Institute of Research Technology & Science, Bhopal, India [3]

ABSTRACT: These days, intrusion detection system (IDS) is the most arising pattern in our general public. This basically screen network traffic and will alarm the organization chairman of any unordinary action. IDS System work by one or the other searching for marks of known assaults or deviations of typical movement. While there are a few detriments of IDS, for example, low recognition rate and high bogus caution rate. Intrusion detection is the process of analyzing the network packets to identify if the packet is legitimate or anomalous. The major challenges involved in this domain includes the huge volume of data for training and the fast and streaming data that is to be provided for the prediction process. Further, the intrinsic data imbalance contained in the domain presents more challenges to the intrusion detection model.

KEYWORDS: IDS, Imbalance, Machine Learning

I. INTRODUCTION

Due to the high demand for internet connectivity between computers, IDS becomes a vital application for network security to inspect various invasion behavior in networks. In our work, we developed network-based IDS which inspects all coming packets and determines any distrustful behavior. In the deployment scenario once malicious traffic is identified, IDS notifies admins/firewalls or intrusion prevention systems. According to methodologies, network-based IDS is divided into two methods which are Signaturedetection and Anomaly-detection. Signature detection works with pre-defined signatures and filters. This technique can inspect the defined intrusions effectively while an undefined attack record is not well determined. In contrast, the anomaly-detection technique relies on heuristic methods to detect the undefined attack behavior. That means when the system identifies a difference from benign traffic patterns then this traffic count as network intrusions. Recent studies showed that a heuristic method is a reliable approach for IDS. The heuristic approach is a proactive and strong method that can use ML and DL algorithms such as unsupervised and supervised. The DL-based system has the ability to extract complex features from a huge amount of network traffic data and can learn the characterization of that complex features. Therefore, the DL based approach can solve the network security tasks and develop a reliable and flexible IDS. The amount of training data and features of data are essential for DL as well. In our work, we used traffic data that was gathered and labeled as either normal or abnormal traffic records from various network resources. DL algorithms can learn reliable feature representations of the data, which can deal with a supervised classification to achieve noticeable outcomes. In our previous work [1] we applied CNN with various architecture and combinations of LSTM and we achieved considerable results. Inspired by our previous work and the promise of DL, we improved and expanded our work by applying various DL models for various datasets. In fact, we used CNN [2,3], LSTM [4], RNN [5,6] and GRU [7] for supervised-learning task with batch normalization and multinomial logistic regression technique. Our deep learning model architectures are depicted in Figures 1–3. We aimed to show the comparison of various DL models evaluation by three different datasets which are UNSW NB15 [8,9], KDDCup '99 [10,11], NSL-KDD [12]. The remainder of the paper is arranged as follows. The Related Works section introduces related works which applied DL algorithms for intrusion detection problems.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

II. LITERATURE REVIEW

Lau lin et al. [1], in imbalanced organization traffic, pernicious digital assaults can regularly stow away in a lot of typical information. It displays a serious level of covertness and jumbling in the internet, making it hard for Network Intrusion Detection System(NIDS) to guarantee the precision and practicality of discovery. This paper explores AI and profound learning for interruption recognition in imbalanced organization traffic. It proposes an original Difficult Set Sampling Technique (DSSTE) calculation to handle the class awkwardness issue. To start with, utilize the Edited Nearest Neighbor(ENN) calculation to separate the imbalanced preparing set into the troublesome set and the simple set. Then, utilize the KMeans calculation to pack the larger part tests in the troublesome set to diminish the larger part. Zoom in and out the minority tests' persistent characteristics in the troublesome set integrate new examples to expand the minority number.

A. Raghavan et al. [2], successful and proficient malware recognition is at the bleeding edge of examination into building secure computerized frameworks. Similarly as with numerous different fields, malware location research has seen a sensational expansion in the utilization of AI calculations. One AI strategy that has been utilized broadly in the field of example matching overall—and malware identification specifically—is covered up Markov models (HMMs). Gee preparing depends on a slope climb, and thus we can frequently work on a model via preparing on numerous occasions with various beginning qualities. In this exploration, we think about helped HMMs (utilizing AdaBoost) to HMMs prepared with different arbitrary restarts, with regards to malware identification. These procedures are applied to an assortment of testing malware datasets. We observe that irregular restarts perform shockingly well in contrast with helping. Just in the most troublesome "cold beginning" situations (where preparing information is seriously restricted) does helping seem to offer adequate improvement to legitimize its higher computational expense in the scoring stage.

Zhiyou Zhang et al. [3], in this paper, aimed at detection of internal intruders in HIDS. Commonly used login ids and passwords may be shared along with co-workers for professional purposes, which can be tampered or used by the attackers as a means of intrusion into the system details. The user was monitored and System Calls (SC) was extracted and the habitual SC pattern based on the habits of the user was taken into account and the profile of the user was stabilized. The forensic technique and other data mining techniques were applied at SC level host IDS to spot the internal attacks. Along with the user login credentials the forensic technique was applied to investigate the computer usage fashion against the collected user profile pattern and thereby check the identity of the user.

Afreen Bhumgara et al. [4], in this paper, With the decision rate threshold of 0.9, the system was able to perform with an accuracy rate of 94%. Nokia Research Center researchers modeled HIDS for mobile devices. The limitation include that each protocol state consume resources for tracing and testing, and its inability to guess the attacks resembling benign protocol. Access control fills in as the cutting edge of resistance against interruptions, bolstering both confidentiality and integrity parameters. Intrusion detection is the process of progressively observing the events occurring in a PC or network, examining them for indications of conceivable episodes and often interdicting the unapproved access. A state transition diagram can be constructed for the sequence of events, but not for the complex forms and hence the attacks having complex behavior which cannot be modeled as the state transition diagram will go unnoticed by the system.

Ritumbhira Uikey et al. [5], in this paper, along with various protection mechanisms accompanied with mobiles they felt an urge for attack monitor methods as a second line of defense. The framework was designed, taking into consideration the privacy of the mobile user in creating the user profile. The framework had a major share with the host-based intrusion detection in line with the network-based detection system, as researchers felt that mobile requires the monitoring system at both ends. The framework included data collection and IDS modules, the former entrusted with responsibility of monitoring the operating system activities, calculating the system measurements and the data collection at the application level and the later feeding on the collected and pre-processed data performs the actual intrusion detection.

Aditya Phadke et al. [6], targeted Advanced Persistent Threat attacks, by analyzing the 30 behavioral pattern of the host user through a 83-dimensional vector, each attribute representing one manner of the user. In order to form the database, they collected 8.7 million features from 4000 malicious and normal programs through the Virtual Machine (VM) environment. The system was designed in such way that frequency of occurrence of each behavior is calculated for each process. C4.5 decision tree was used to build a classifier for the collected information, and each new instance was analyzed

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

against the tree to be segregated as malicious or normal instance. The model had a false positive rate of 5.8% and a false negative rate of 2.0%.

S. Sivantham et al. [7], in this paper, represented a novel HIDS aimed at discovering unknown malware codes. The collection of previous malware codes was taken as repository and each new sequence of behavior was compared with the repository to identify new malware code. Applied rule-based IDS to tackle the DDoS attacks in which the resources are made unavailable for the user when they are required. The utmost capacity of each of the middle-ware layer was fixed and set of rules were formed to detect the DDoS attacks. The system produced an alert when the count of the requests to a particular resource exceeded a particular threshold and concepts from learning automata were employed to avoid further attacks.

Azar Abid Salih et al. [8], in this paper, applied Machine Learning techniques, namely Naive Bayes, a Bayesian Network and Artificial Neural Network, to perform supervised learning of the malicious code. They gathered 323 features for training the classifiers. The detection rate for a specific set of worms was over 98%. Though HIDS were able to perform better by centering the user profile collected across, it prompted the challenge and there was a lack of information of the user Centralized reporting wasn't feasible with HIDS. They consumed the host details and resources which may violate the privacy issues of the user and may dispute with the already existing security protocols.

Lukman Hakim et al. [9], in this paper, centered over detecting intrusion in Routing Protocol for Low Power and Lossy Networks (RPL) attacks. The operations of the RPL were converted into finite state machines through which the network was monitored and any malicious activity was detected. The research was further extended by wherein the simulation trace files were used to model the finite state machines to observe the RPL attacks. The model was further converted into a set of rules to monitor the data transferred between the network nodes. The drawback of the work was that the True Positive Rate was even able to reach 100% but the False Positive Rate was not that low ranging between 0 to 6.78%. Over that it also had an overhead of 6.3% in terms of energy when compared to normal RPL network.

T. Sree Kala et al. [10], in this paper, designed an IDS for cloud environment based on state transition methodology. A Hidden Markov Model which builds a model where the behavior of user observed over long time period is marked as states and relevant transitions between them was used in their research. They developed three profiles, namely, low, middle and high, based on the matching of the probability of the user to the baseline profile. They give input as VM SC and bring about a diagram of state transition, wherein each transition represents the probability of targeting the next state and the probability of creating next system call. This model is tested against DoS attacks with 100% detection rate but with a poor false positive rate of 5.66%.

III. INTRUSION DETECTION SYSTEM

An Intrusion Detection System (IDS) is a monitoring system that detects suspicious activities and generates alerts when they are detected. Based upon these alerts, a security operations center (SOC) analyst or incident responder can investigate the issue and take the appropriate actions to remediate the threat.

Classification of Intrusion Detection Systems

Intrusion detection systems are designed to be deployed in different environments. And like many cybersecurity solutions, an IDS can either be host-based or network-based.

Host-Based IDS (HIDS): A host-based IDS is deployed on a particular endpoint and designed to protect it against internal and external threats. Such an IDS may have the ability to monitor network traffic to and from the machine, observe running processes, and inspect the system's logs. A host-based IDS's visibility is limited to its host machine, decreasing the available context for decision-making, but has deep visibility into the host computer's internals.

Network-Based IDS (NIDS): A network-based IDS solution is designed to monitor an entire protected network. It has visibility into all traffic flowing through the network and makes determinations based upon packet metadata and contents.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

This wider viewpoint provides more context and the ability to detect widespread threats; however, these systems lack visibility into the internals of the endpoints that they protect.
Due to the different levels of visibility, deploying a HIDS or NIDS in isolation provides incomplete protection to an organization's system. A unified threat management solution, which integrates multiple technologies in one system, can provide more comprehensive security.

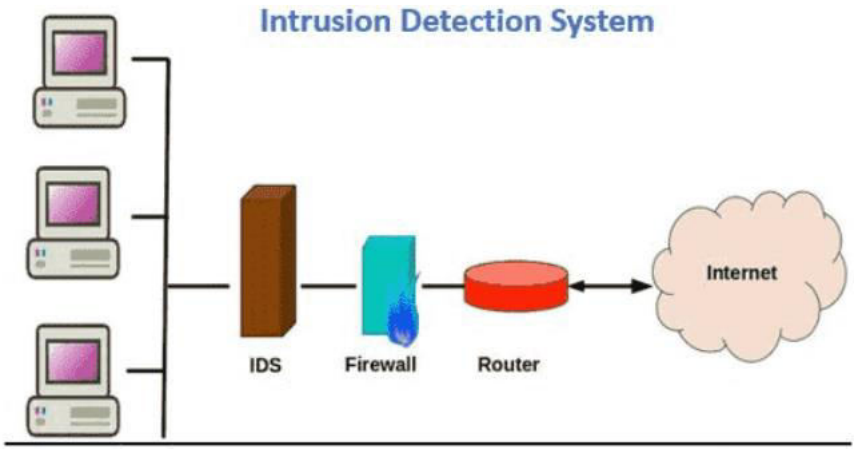

Fig. 1: IDS System

IV. ML ALGORITHM

Supervised learning is two stage forms, in the initial step: a model is fabricate depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset.

Learning
The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [9].

- Supervised learning
- Unsupervised learning

Supervised Learning
Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset.

Unsupervised learning
It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.
Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set.

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 1.4. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error. Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations.

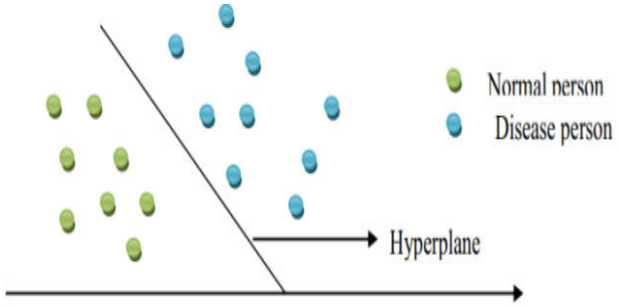

Fig. 2: Support Vector Classification

V. METHODOLOGY

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

LSTM has a chain structure that contains four neural networks and different memory blocks called **cells**.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

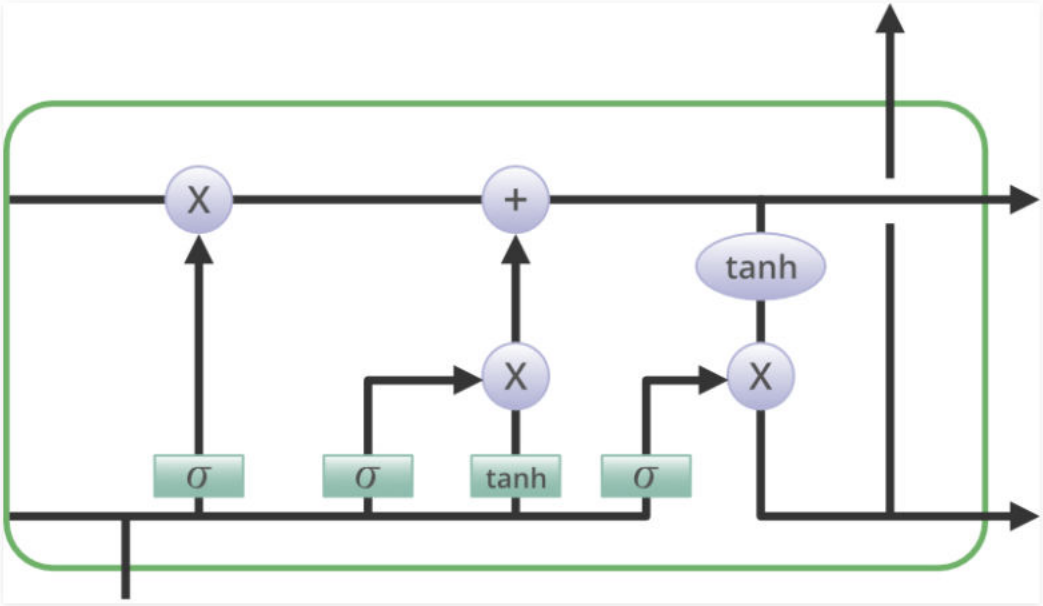

Fig. 3: LSTM

Information is retained by the cells and the memory manipulations are done by the **gates.** There are three gates –

1. Forget Gate: The information that is no longer useful in the cell state is removed with the forget gate. Two inputs *x_t* (input at the particular time) and *h_t-1* (previous cell output) are fed to the gate and multiplied with weight matrices followed by the addition of bias. The resultant is passed through an activation function which gives a binary output. If for a particular cell state the output is 0, the piece of information is forgotten and for output 1, the informati on is retained for future use.
2. Input gate: The addition of useful information to the cell state is done by the input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate using inputs *h_t-1* and *x_t*. Then, a vector is created using *tanh* function that gives an output from -1 to +1, which contains all the possible values from h_t-1 and *x_t*. At last, the values of the vector and the regulated values are multiplied to obtain the useful information
3. Output gate: The task of extracting useful information from the current cell state to be presented as output is done by the output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter by the values to be remembered using inputs *h_t-1* and *x_t*. At last, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.

VI. CONCLUSION

In the previous few decades, internet technology has extended its application space in many various domains in our life like banking operations, on-line auctions, electronic commerce applications, social networking, and on-line application / registration, etc. However, because of the weakness of computer systems' security, numerous electronic networks have typically been intruded by the hackers particularly with denial-of-service or distributed denial-of- service attacks. Anomaly detection principally depends on the illustration of the conventional behavior of the users, hosts and network connections. It is extremely laborious to notice abnormal requests within the network. Therefore, machine learning algorithms are used as a versatile and powerful technique.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 11, Issue 3, March 2022 ||

| DOI:10.15680/IJIRSET.2022.1103098

REFERENCES

[1] Lan Liu, Pengcheng Wang , Jun Lin, and Langzhou Liu, “Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning”, IEEE Access 2020.

[2] A. Raghavan, F. D. Troia, and M. Stamp, ``Hidden Markov models with random restarts versus boosting for malware detection," *J. Comput. Virol. Hacking Techn.*, vol. 15, no. 2, pp. 97107, Jun. 2019.

[3] Zhiyou Zhang and Peishang Pan “A hybrid intrusion detection method based on improved fuzzy C-Means and SVM”, IEEE International Conference on Communication Information System and Computer Engineer (CISCE), pp. no. 210-214, Haikou, China 2019.

[4] Afreen Bhumgara and Anand Pitale, “Detection of Network Intrusion Using Hybrid Intelligent System”, IEEE International Conferences on Advances in Information Technology, pp. no. 167-172, Chikmagalur, India 2019.

[5] Ritumbhira Uikey and Dr. Manari Cyanchandani “Survey on Classification Techniques Applied to Intrusion Detection System and its Comparative Analysis”, IEEE 4th International Conference on Communication $ Electronics System (ICCES), pp. no. 459-466, Coimbatore, India 2019.

[6] Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar and Rashmi Bhattad “A Review of Machine Learning Methodologies for Network Intrusion Detection”, IEEE 3rd National Conference on Computing Methodologies and Communication (ICCMC), pp. no. 703-709, Erode, India 2019.

[7] S. Sivantham, R.Abirami and R.Gowsalya “Comapring in Anomaly Based Intrusion Detection System for Networks”, IEEE International conference on Vision towards Emerging Trends in Communication and Networking (ViTECon), pp. no. 289-293, Coimbatore, India 2019.

[8] Azar Abid Salih and Maiwan Bahjat Abdulrazaq “Combining Best Features selection Using Three Classifiers in Intrusion Detection System”, IEEE International Conference on Advanced science and Engineering (ICOASE), pp. no. 453-459, Zakho - Duhok, Iraq 2019.

[9] Lukman Hakim and Rahilla Fatma Novriandi “Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset”, IEEE International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp. no. 330-336, Jember, Indonesia 2019.

[10] T. Sree Kala and A. Christy, “An Intrusion Detection System Using Opposition Based Particle Swayam Optimization Algorithm and PNN”, IEEE International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp. no. 564-569, Coimbatore, India 2019.

[11] Xiaoyan Wang and Hanwen Wang “A High Performance Intrusion Detection Method Based on Combining Supervised and Unsupervised Learning”, IEEE Smart World, Ubiquitous Intelligence $ Computing Advanced $ Trusted Computing, Scalable Computing, Internet of People and Smart City Innovations, pp. no. 889-897, Guangzhou, China 2018.

[12] P. Singh and M. Venkatesan, “Hybrid Approach for Intrusion Detection System”, IEEE International Conference on Current Trends Towards Converging Technologies (ICCTCT), pp. no. 654-659, Coimbatore, India 2018.

[13] M. Tavallaee, E. Bagheri, W, Lu and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 dataset”, IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. no. 892-899, Ottawa, India 2018.

[14] Karuna S. Bhosale and Assoc. prof. Maria, “Data Mining Based Advanced Algorithm for Intrusion Detection in Communication Networks”, IEEE International Conference on Computational Techniques, Electronics & Mechanical System (CTEMS), Belgaum, India 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details