Designing of Intrusion Detection System using Data Mining Techniques

[1] Mrs. R. Chithra, [2] M. Lakshmi Priya

[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract: Due to the widespread diffusion of network connectivity, the demand for network security and protection against cyber-attacks is ever increasing. Intrusion detection systems (IDS) perform an essential role in today's network security. This paper proposes an IDS based on feature selection and clustering algorithm using filter and wrapper methods. Filter and wrapper methods are named feature grouping based on linear correlation coefficient (FGLCC) algorithm and cuttlefish algorithm (CFA), respectively. Artificial Neural Network is used as the classifier in the proposed method. For performance verification, the proposed method was applied on KDD Cup 99 large data sets. The results verified a high accuracy (95.03%) and detection rate (95.23%) with a low false positive rate (1.65%) compared to the existing methods in the literature**.**

***Index Terms* - Cyber Security, Intrusion Detection System, Feature Selection, Classification, CuttleFish algorithm, Linear Correlation Coefficient, Artificial Neural Network.**

I. INTRODUCTION

With the rapid expansion of the computer network during the past few years, the information security issue becomes more and more important. There are many research topics for network security. Like as, data encryption, vulnerability database, intrusion detection, etc. Intrusion detection is one of the major information security problems [1][2]. IDS (Intrusion Detection System) assist the system in resisting external attacks. Existing IDS can be divided into two categories according to the detection approaches: anomaly detection and misuse detection or signature detection. Data mining techniques can be used for misuse and anomaly intrusion detection [3][4]. Misuse refers to known attacks and harmful activities that exploit the known sensitivities of the system. In misuse detection, each instance in a data set is labeled as ―normal‖ or‖ intrusion‖ and a learning algorithm is trained over the labeled data. Anomaly means a usual activity in general that could indicate an intrusion. An advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variation [5]. As there are many number of ID techniques using data mining techniques, the unknown technique and system could be thought of as a baseline for future prospect. As a result, the purpose of this paper is to review related papers of using data mining for intrusion detection. The contribution of this research paper is to provide a comparison of IDS in terms of data mining IDS techniques used for future research directions [6][7].

An Intrusion Detection System is used to detect all types of malicious network traffic and computer usage that cannot be detected by an ordinary firewall. It includes network attacks against sensitive services, data driven attacks on computer applications, host based attacks such as privilege and permissions escalation, unauthorized logins and access to sensitive files, and malware (viruses, Trojan horses, and worms) [8][9]. IDS are the best fine grain filter placed inside the protected computer network, looking for known or powerful threats in network traffic and/or audit data recorded by hosts [10].

II. RELATED WORKS

Khraisat, Ansam, Iqbal Gondal, and Peter Vamplew [11] This paper examines different data mining techniques that could minimize both the number of false negatives and false positives. C5 classifier's effectiveness is examined and compared with other classifiers. Results should that false negatives are reduced and intrusion detection has been improved significantly. A consequence of minimizing the false positives has resulted in reduction in the amount of the false alerts as well. In this study, multiple classifiers have been compared with C5 decision tree classifier using NSL_KDD dataset and results have shown that C5 has achieved high accuracy and low false alarms as an intrusion detection system.

Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein [12] A hybrid approach with two main parts is proposed to address these issues. First, data needs to be filtered using the Vote algorithm with Information Gain that combines the probability distributions of these base learners in order to select the important features that positively affect the accuracy of the proposed model. Next, the hybrid algorithm consists of following classifiers: J48, Meta Pagging, RandomTree, REPTree, AdaBoostM1, DecisionStump and NaiveBayes. Based on the results obtained using the proposed model, we observe improved accuracy, high false negative rate, and low false positive rule.

Sharma, Ruby, and Sandeep Chaurasia [13] This paper proposes an Intrusion Detection System based on the density maximization-based fuzzy c-means clustering (DM-FCC). In this approach, cluster efficiency is improved through a membership matrix generation (MMG) algorithm. Dissimilarity Distance Function (DDF) has been used to compute the distance metric while creating a cluster in proposing an IDS. The proposed enhanced fuzzy c-means algorithm has been tested upon ADFA Dataset and the model performs highly appreciable in terms of accuracy, precision, detection rates, and false alarms.

Saxena, Akash, Khushboo Saxena, and Jayanti Goyal [14] In our proposed work, we initially apply KDD cup'99 dataset which is most broadly used method for detecting intrusion. DBSCAN is the most utilized method which is used to eliminate noise from the data. Then, we generate the most meaning inputs by analyzing and processing whole data which is done by the selection of feature method. K-means clustering performs grouping of data which is followed by SMO classifier. So we proposed a hybrid structure which improves the taken as a whole accuracy. MATLAB and WEKA tools are used to execute the whole process.

Gupta, Amara SALG Gopal, G. Syam Prasad, and Soumya Ranjan Nayak [15] All through this algorithmic program, intruder only ready to recoup or improve key by communicating with the Intrusion Detection System and perspective the tip result after it and by abuse this theme can't prepared to meet security norms. In this way supported learning we'd quite recently like the topic that can assist us with providing extra security on Data Storage. To reduce the attack risk, a dynamic key theory is bestowed and analyzed we've an inclination to face live about to planned theme for extra security that is ready to be secure delicate information of fluctuated domains like in consideration area enduring associated information like contact points of interest and antiquity.

III. PROBLEM STATEMENT

Anomaly-based detection techniques rely on the assumption that the intruder's behavior is different from normal network behaviours. These techniques study the normal traffic of the net- work and identify each deviant behavior as malicious behavior. This system provides the possibility of detecting both unknown and known attacks. The main disadvantage of this system is its high false positive rate[19][20][21][22][23].

- The reduced classification accuracy in the intrusion detection system.
- Increased error rates like Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE)
- Decreased precision and increased false positive rate.

IV. PROPOSED METHODOLOGY

In statistics, the correlation coefficient indicates the strength and direction of a relationship between two random variables. The commonest use refers to a linear relationship. In general, statistical usage, correlation or correlation refers to the departure of two random variables from independence. Equation (1) shows the calculation of the correlation coefficient between two variables x and y. There are totally n observations.

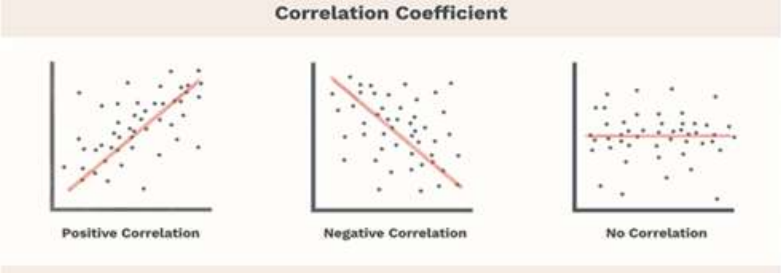

Figure 1: Types of Relationship in Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\ \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Two variables have strong dependency when their correlation coefficient value is close to 1 or -1. When the value is 0, it means that the two variables are not related at all. In our research, strong dependency is what we are looking for, no matter it is positive or negative. Therefore, in the measurement procedure, the absolute value of the correlation coefficient | r | is used.

4.1Cuttlefish Algorithm

A new meta-heuristic bio-inspired optimization algorithm, called Cuttlefish Algorithm (CFA) is presented. The algorithm mimics the mechanism of color changing behavior used by the cuttlefish to solve numerical global optimization problems. The patterns and colors seen in cuttlefish are produced by reflected light from different layers of cells including (chromatophores, leucophores and iridophores) stacked together, and it is the combination of certain cells at once that allows cuttlefish to possess such a large array of patterns and colors. The proposed algorithm considers two main processes: reflection and visibility. Reflection process is proposed to simulate the light reflection mechanism used by these three layers, while the visibility is proposed to simulate the visibility of matching pattern used by the cuttlefish. These two processes are used as a search strategy to find the global optimal solutions.

4.2 Artificial Neural Network

Artificial Neural Network (ANN) is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. ANNs are also named as "artificial neural systems," or "parallel distributed processing systems," or "connectionist systems." ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. In order to form a feed-forward multi-layer in MLP, the collection of non-linear neurons is connected to one another. This technique is known to be very useful for prediction and classification issues. Cross-validation is used to determine the 'optimal' number of hidden layers and neurons which were relied on the experimental design of the Intrusion Detection classification framework. These units, also referred to as nodes or neurons, are simple processors which operate in parallel[16][17][18].

V. RESULTS AND DISCUSSION

5.1 Dataset Description

With the widespread use of computer networks, the number of attacks has grown extensively, and many new hacking tools and intrusive methods have appeared. Using an intrusion detection system (IDS) is one way of dealing with suspicious activities within a network. An intrusion detection system (IDS) monitors networked devices and looks for anomalous or malicious behavior in the patterns of activity in the audit stream. There are two general types of Intrusion Detection systems, they are Host based Intrusion Detection System and Network based Intrusion Detection System. The Host based Intrusion Detection system has host based sensors and the network based Intrusion detection system has network-based sensors. Data mining-based intrusion detection systems can be classified according to their detection strategy. There are two main strategies: misuse detection, which uses patterns of well-known attacks or weak spots of the system to identify intrusions and anomaly detection, which tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions. (a)Misuse Detection: Misuse detection attempts to model abnormal behavior based on signatures of the known attacks and known system vulnerabilities. (b) Anomaly Detection: Normal behavior patterns are useful in predicting both user and system behavior. Anomaly detectors construct profiles that represent normal usage and then use current behavior data to detect a possible mismatch between profiles and recognize possible attack attempts. Signature-based schemes provide very good detection results for specified, well-known attacks. However, they are not capable of detecting new, unfamiliar intrusions, even if they are built as minimum variants of already known attacks. On the contrary, the main benefit of anomaly-based detection techniques is their potential to detect previously unseen intrusion events.

KDD training dataset consists of relatively 4,900,000 single connection vectors where each single connection vectors consists of 41 features and is marked as either normal or an attack, with exactly one particular attack type. These features had all forms of continuous and symbolic with extensively varying ranges falling in four categories:

- In a connection, the first category consists of the intrinsic features which comprises of the fundamental features of each individual TCP connections. Some of the features for each individual TCP connections are duration of the connection, the type of the protocol (TCP, UDP, etc.) and network service (http, telnet, etc.)
- The content features suggested by domain knowledge are used to assess the payload of the original TCP packets, such as the number of failed login attempts.
- Within a connection, the same host features observe the recognized connections that have the same destination host as present connection in past two seconds and the statistics related to the protocol behaviour, service, etc are estimated.
- The similar same service features scrutinize the connections that have the same service as the current connection in past two seconds.

Table 1: Attacks falling into four major categories

Category	Number of Attacks
Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
User to Root Attacks	Buffer_overflow, loadmodule, perl, rootkit,
Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
Probes	Satan, ipsweep, nmap, portsweep

Table 2: Features given in KDD cup 99 dataset

Feature index	Feature name	Description	Type
1	duration	length (number of seconds) of the connection	Continuous
2	protocol_type	type of the protocol, e.g. tcp, udp,icmp etc.	Symbolic
3	Service	network service on the destination, e.g., http, telnet, etc.	Symbolic
4	Flag	normal or error status of the connection	Symbolic
5	src_bytes	number of data bytes from source to destination	Continuous
6	dst_bytes	number of data bytes from destination to source	Continuous
7	Land	1 if connection is from/to the same host/port; 0 otherwise	Symbolic
8	wrong_fragment	number of ``wrong" fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	number of ``hot" indicators	Continuous
11	num_failed_logins	number of failed login attempts	Continuous
12	logged_in	1 if successfully logged in; 0 otherwise	Symbolic
13	num_compromised	number of ``compromised" conditions	Continuous
14	root_shell	1 if root shell is obtained; 0 otherwise	Continuous
15	su_attempted	1 if ``su root" command attempted; 0 otherwise	Continuous
16	num_root	number of “root” accesses	Continuous
17	num_file_creations	number of file creation operations	Continuous
18	num_shells	number of shell prompts	Continuous
19	num_access_files	number of operations on access control files	Continuous

20	num_outbound_cmds	number of outbound commands in an ftp session	Continuous
21	is_hot_login	1 if the login belongs to the ``hot" list; 0 otherwise	Symbolic
22	is_guest_login	1 if the login is a ``guest" login; 0 otherwise	Symbolic
23	Count	number of connections to the same host as the current connection in the past two seconds	Continuous
24	srv_count	number of connections to the same service as the current connection in the past two seconds	Continuous
25	serror_rate	% of connections that have ``SYN" Errors	Continuous
26	srv_serror_rate	% of connections that have ``SYN" Errors	Continuous
27	rerror_rate	% of connections that have ``REJ" Errors	Continuous
28	srv_rerror_rate	% of connections that have ``REJ" Errors	Continuous
29	same_srv_rate	% of connections to the same service	Continuous
30	diff_srv_rate	% of connections to different Services	Continuous
31	srv_diff_host_rate	% of connections to different hosts	Continuous
32	dst_host_count	count for destination host	Continuous
33	dst_host_srv_count	srv_count for destination host	Continuous
34	dst_host_same_srv_rate	same_srv_rate for destination host	Continuous
35	dst_host_diff_srv_rate	diff_srv_rate for destination host-	Continuous
36	dst_host_same_src_port_rate	same_src_port_rate for destination host	Continuous
37	dst_host_srv_diff_host_rate	diff_host_rate for destination host	Continuous
38	dst_host_serror_rate	serror_rate for destination host	Continuous
39	dst_host_srv_serror_rate	srv_serror_rate for destination host	Continuous
40	dst_host_rerror_rate	rerror_rate for destination host	Continuous
41	dst_host_srv_rerror_rate	srv_serror_rate for destination host	Continuous

5.2 Number of Features obtained by proposed Feature Selection Method

The following table 3 depicts the number of features obtained by using Ant Colony Optimization, Particle Swarm Optimization and Linear Correlation Coefficient feature selection methods. The following 27features are obtained by correlation coefficient feature selection method.

Table 3: Number of Features obtained by Linear correlation Coefficient Feature Selection Method

Sl.No	Ant Colony Optimization	Particle Swarm Optimization	Linear Correlation Coefficient
1	Src_bytes	logged_in	Service
2	Service	num_5ell	srv_Count
3	srv_Count	dst_host_same_src_port_rate	count
4	count	dst_host_diff_srv_rate	dst_host_srv_count
5	dst_host_srv_count	Protocol_type	dst_host_same_src_port_rate
6	dst_host_same_src_port_rate	dst_host_same_srv_rate	dst_host_diff_srv_rate
7	Dst_bytes	dst_host_srv_serror_rate	dst_host_same_srv_rate
8	dst_host_diff_srv_rate	num_access_fil	dst_host_rerror_rate
9	dst_host_same_srv_rate	srv_Count	dst_host_srv_rerror_rate
10	dst_host_rerror_rate	num_compromised	dst_host_srv_diff_host_rate
11	dst_host_srv_rerror_rate	num_root	diff_srv_rate
12	dst_host_count	srv_serror_rate	Flag
13	dst_host_srv_diff_host_rate	rerror_rate	same_srv_rate
14	diff_srv_rate	srv_diff_host_rate	srv_diff_host_rate
15	Flag	hot	rerror_rate
16	same_srv_rate	dst_host_srv_diff_host_rate	srv_rerror_rate
17	srv_diff_host_rate	is_guest_32	dst_host_serror_rate
18	rerror_rate	dst_host_srv_count	logged_in
19	srv_rerror_rate	Flag	dst_host_srv_serror_rate
20	dst_host_serror_rate	su_attempted	serror_rate
21	logged_in	Root_5ell	srv_serror_rate
22	duration	dst_host_rerror_rate	hot
23	dst_host_srv_serror_rate	Wrong_fragment	Wrong_fragment
24	serror_rate	diff_srv_rate	num_failed_32s
25	srv_serror_rate	serror_rate	num_compromised
26	hot	dst_host_serror_rate	is_guest_32
27	Wrong_fragment	srv_rerror_rate	land
28	num_failed_32s	num_file_creations	
29	num_compromised	dst_host_srv_rerror_rate	
30	is_guest_32		
31	land		

5.3 Performance Analysis of the feature selection method using classification methods

Mostly the classification methods are utilized to evaluate the effectiveness of the results obtained from the feature selection techniques. In this research work, the classification methods like ANN and NB have used. By using ANN classification method. Table 4 depicts the Linear Correlation Coefficient method gives better classification accuracy, Kappa Statistic value, reduced error rates like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE) and Relative Root Absolute Error (RRAE) than the existing methods like Chi-Square Analysis and Particle Swarm Optimization.

Table 4: Performance analysis of the original dataset, ACO, PSO and Linear correlation Coefficient feature selection by using Naïve Bayes Classification Method

Performance Metrics	Original Dataset	*Feature Selection Methods*		
		ACO	PSO	Linear Correlation Coefficient
Classification Accuracy	65.3333 %	75.3333 %	71.3333 %	92.6667 %
Kappa Statistic	0.4664	0.5083	0.4935	0.7432
MAE	0.0198	0.0292	0.0225	0.0128
RMSE	0.1354	0.157	0.1448	0.1077
RAE	44.2956 %	42.273 %	39.4545 %	20.0334 %
RRSE	92.471 %	87.1663 %	88.0308 %	64.0312 %
TP Rate	0.653	0.753	0.713	0.927
FP Rate	0.22	0.26	0.247	0.251
Precision	0.494	0.647	0.549	0.865
Recall	0.653	0.753	0.713	0.927
F-Measure	0.545	0.691	0.619	0.894
ROC Area	0.8	0.822	0.859	0.943

Table 5: Performance analysis of ACO, PSO and Linear Correlation Coefficient Feature Selection method by using Artificial Neural Network

Performance Metrics	Original Dataset	*Feature Selection Methods*		
		ACO	PSO	Linear Correlation Coefficient
Classification Accuracy	69.3333 %	94.6667 %	92 %	98 %
Kappa Statistic	0.5539	0.7133	0.7099	0.7594
MAE	0.0411	0.0934	0.0874	0.0788
RMSE	0.1317	0.1671	0.1326	0.1311
RAE	91.7907 %	96.1844 %	90.0778 %	73.7135 %
RRSE	89.9791 %	78.975 %	78.8312 %	60.0193 %
TP Rate	0.693	0.947	0.92	0.98
FP Rate	0.15	0.294	0.307	0.355
Precision	0.553	0.923	0.853	0.98
Recall	0.693	0.947	0.92	0.98
F-Measure	0.601	0.933	0.885	0.978
ROC Area	0.847	0.859	0.93	1

V. CONCLUSION

The pre-processing method has implemented to excrete the redundant and irrelevant features from the dataset. This proposed technique has used to improve the prognostication accuracy. In this work, a Linear Correlation Coefficient feature selection method has introduced by comparing the PSO and ACO analysis. This method was presented to eliminate the unnecessary feature for the classification in the IDS dataset. From the obtained results it has determined that the proposed technique worked better than the present feature selection method in the IDS. Also, it enhances the prognostication accuracy and diminishes the error rates. This diminishing of error rates results in the excellent classification accuracy.

With the reduced dataset, the classification accuracy has increased by using the ANN classification method. The error rates are reduced and true positive, ROC values have increased. By using this methodology, the node can be classified into a malicious and legitimate category.

REFERENCES

[1] R. Akbani, Korkmaz, T., Raju, G. V. S: Mobile ad hoc network security. In: Lecture Notes in Electrical Engineering, Springer, vol. 127 (2012)

[2] T. Anantvalee and Wu, J.: A survey on intrusion detection in mobile ad hoc networks. In: Wireless/Mobile Security. New York: Springer (2008)

[3] Elhadi, M., Shakshuki, EAACK.: A secure intrusion-detection system for MANETs. In: IEEE Transactions on Industrial Electronics, vol. 60(3) (2013)

[4] Gungor, V.C., Hancke, G.P.: Industrial wireless sensor networks: challenges, design principles, and technical approach. IEEE Trans. Ind. Electron. 56(10), 4258–4265 (2009)

[5] Haldar, N.A.H.: An activity pattern based wireless intrusion detection system. In: Information Technology: pp. 846–847 (2012)

[6] Shen, J.: Network intrusion detection by artificial immune system, IECON, pp. 716–720 (2011)

[7] Khattab, S., Gabriel, S., Melhem, R., Mosse, D.: Live baiting for service-level DoS attackers. Proceeding of the IEEE INFOCOM (2008)

[8] Macia´-Pe´rez, F.: Network intrusion detection system embedded on a smart sensor, industrial electronics. IEEE Trans. 58(3), 722– 732 (2012).

[9] Benjie Chen, Kyle Jamieson, Hari Balakrishnan And Robert Morris" An Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks," in Proceedings of the wireless network, 2002.

[10] Krishan Kumar, "power control with transition time for the ad-hoc wireless network," International Journal of Advanced Research in Computer Science, Volume 3, No. 3, May-June 2012.

[11] Khraisat, Ansam, Iqbal Gondal, and Peter Vamplew. "An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2018

[12] Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model." *Journal of Computational Science* 25 (2018): 152-160

[13] Sharma, Ruby, and Sandeep Chaurasia. "An enhanced approach to fuzzy C-means clustering for anomaly detection." *Proceedings of First International Conference on Smart System, Innovations and Computing*. Springer, Singapore, 2018

[14] Saxena, Akash, Khushboo Saxena, and Jayanti Goyal. "Hybrid Technique Based on DBSCAN for Selection of Improved Features for Intrusion Detection System." *Emerging Trends in Expert Applications and Security*. Springer, Singapore, 2019. 365-377

[15] Gupta, Amara SALG Gopal, G. Syam Prasad, and Soumya Ranjan Nayak. "A New and Secure Intrusion Detecting System for Detection of Anomalies Within the Big Data." *Cloud Computing for Geospatial Big Data Analytics*. Springer, Cham, 2019. 177-190.

[16] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems,* Preprint: 1-18.

[17] M. Durairaj, T S Poornappriya, "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.", *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[18] M. Durairaj, T. S. Poornappriya, "Importance of MapReduce for Big Data Applications: A Survey", *Asian Journal of Computer Science and Technology,* Vol.7 No.1, 2018, pp. 112-118.

[19] M. Lalli, V.Palanisamy,(2016), "Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks", International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[20] M. Lalli, V.Palanisamy,(2017), "Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques", Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[21] M. Lalli, V.Palanisamy, (Sep 2014), "A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN", International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[22] M. Lalli, "Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET", *Journal of Applied Science and Computations,* Volume VI, Issue VI, JUNE/2019, 1795-1813.

[23] M. Lalli, "Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network", *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.

.