

A Hybrid Approach for User to Root and Remote to Local Attack

R. Richhariya^{1*}, A.K. Manjhar², R. R. Singh Makwana³

^{1*} Dept. of CSE/IT, Madhav Institute of Technology and Science, Gwalior, India

² Dept. of CSE /IT, Madhav Institute of Technology and Science, Gwalior, India

³ Dept. of CSE /IT, Madhav Institute of Technology and Science, Gwalior, India

**Corresponding Author: yadakada@gmail.com*

Available online at: www.ijcseonline.org

Received: 23/May/2017, Revised: 06/Jun/2017, Accepted: 26/Jun/2017, Published: 30/Jun/2017

Abstract- With the monstrous grow in the usage of computers and Internet for information sharing, success of applications running on various platforms causes a serious risk to the security policy. Complex behavior of malwares are also increase, the mechanism to catch malwares also needs improvement. The challenges grow towards the network security due to the introduction of new attacks. This paper emphasize on a hybrid data-mining approach based on ensemble classifier. This preferred approach gives a hybrid classifier which improves the overall detection rate. Preferred approach gives more accuracy and decrease the false positive rate. With this preferred approach the classification accuracy is 99.9894% and the false positive rate is about 0.00. The comparison of preferred approach is made with the single best classifier and it is perceived that the preferred approach gives better results for User to Root (U2R) and Remote to Local (R2L) attack that present in NSL KDD intrusion dataset. This approach gives better results for Root-kit attack.

Keywords- Intrusion detection system (IDS), false positive rate, NSL-KDD dataset

I. INTRODUCTION

Data-mining take the important role in IDS. IDS using data-mining are mainly used due to the excellence performance in accuracy, classification and strong responsive nature as the new ever changing difficult environment. In Intrusion Detection the data divided into two main streams which help to arrange the data into more simplified manner. IDS safeguards the systems from unauthorized users, it diversifies the intrusion and categorizes into “bad” connections i.e. attacks and “good” i.e. normal connections. An IDS finds out the actual reason for example (attacks for the outer resource) and malfunctioning (attacks within the organization).

In intrusion detection the key thought of applying data-mining methods is computerization. Machine learning is the capability of a machine that consequently enhances its performance completely through learning from experience. Data-mining methods such as decision trees, neural networks, support vector machine, naïve Bayesian, bayes net classifiers and many other algorithms have been utilize to determine the network logs and to catch intrusion related information to get better correctness of IDS.

The process of KDD involves various steps that is interactive, friendly, and iterative and decision making rules which is user-driven. Data-mining is the very important part in the KDD process, and it utilizes data-mining approaches to extract patterns within the data. The functionalities of data-mining is utilize to report the various patterns which is to found in the tasks of data-mining and it can be divided into two categories:

- Descriptive: which is to report the data consist of general properties in the dataset
- Predictive: this is used for predictions as well as to perform inference on items [1].

II. INTRUSION DETECTION SYSTEM

Intrusion detection system (IDS) is a combination of software and hardware that monitor and examines the actions occurring in network system to distinguish intrusions. Intrusion detection using data-mining plays a Key role of analyzing, processing and sorting the data to a systematic and organized manner without any errors. Traditional IDS search for possible malicious events on network traffics and occasionally finds real security attacks and abnormalities. But, many times fails to distinguish malicious actions or identify normal traffic as a attack type in the network. To enhance the capability of intrusion detection system, Data mining methods are applied on network data because of it can process great volume of data and User’s subjective evaluation is not necessary, and it is more suitable to discover the ignored and hidden information. There are two types of IDS [2]:

- Signature Based IDS (SBIDS) and
- Anomaly Based IDS (ABIDS).

III. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

A dataset is unbalanced if the classification classes don't seem to be some uniformly distributed. The real-world knowledge sets are mainly composed of "normal" examples with solely a little proportion of "attack" examples. It is conjointly the case that the value of misclassifying associate degree attack example as a traditional example is commonly abundant more than the utility of the reverse error. The efficiency of machine learning algorithms is often evaluated predictive accuracy. However, this is often not applicable once the information is unbalanced and/or the costs of various errors vary markedly.

There are two techniques of re-sampling that are employed for increasing the sensitivity of a classifier to the minority class: Under-sampling of the majority (legitimate) class and Over-sampling the minority (illegitimate) class. To address the unbalanced dataset problem, in this paper we used Synthetic Minority Oversampling Technique (SMOTE) as a preprocessor. In SMOTE artificial samples square measure generated on the road segments change of integrity the k minority category nearest neighbors. The number of oversampling in SMOTE is influenced by the quantity of randomly selected neighbors from the k -nearest neighbors. SMOTE generates artificial samples by taking the distinction between feature vector of the instance into deliberation and its nearest neighbor then multiplying this distinction by a random number between zero and one, then add this product to the feature vector into consideration [3].

IV. FEATURE SELECTION

Here variable selection is also known as "Feature selection, or vice a versa, this process commonly used within machine learning, where some sets of the features are obtainable from the dataset is chosen for learning algorithm's application. This process is necessary in two situations first because it's infeasible computationally where all available features are to be used, or due to estimation problems, when limited data samples (but a more number of features) are present". Feature selection from the obtainable data is vital to the effectiveness of the methods employed. Researchers use different analysis approaches in a process till the accumulation of data, in order to select the set of features that they think increases the effectiveness of their techniques called data mining. Each of these selected features offers a valuable piece of information to the System. The features which are extracted are ranked due their whole contribution and are accordingly utilized.

V. RANDOM FOREST

Random forest classification algorithmic rule uses ensemble ways to get higher predictive performance. It produces output within the type of individual trees and relies upon the decision tree algorithmic rule. It's thought about to be an extremely correct classifier and may handle multiple

variables [4]. The algorithm used for receiving a random forest that was developed by the developer Adele monger, and Leo Breiman and "Random Forests" is their trademark. The term "Random Forests" came from random call forest that was projected first of all in 1995 by Tin Kam metal of Bell Labs. The strategy combines two ideas that are of Breiman's "bagging" plan and also the random choice of options that is introduced separately by Ho, Amit and Geman for the development of assortment of call trees comprises controlled variation.

VI. RANDOM TREE

It generates a tree by randomly selecting branches from a possible set of trees. The trees are distributed in a uniform way so chances of getting sampled are equiprobable.

VII. BAYES NET

Bayes net is a widely used technique which works on the basic Bayes theorem and forms a Bayesian network after calculating conditional probability on each node. It is a pictorial model which is probabilistic in nature and portrays a group of arbitrary variables along with their conditional dependencies through a directed acyclic graph.

VIII. DATASET DESCRIPTION

The NSL-KDD dataset is the dataset that is openly available for intrusion detection research. It provides labels for both training and test data sets. The dataset was created based on the 1998 DARPA intrusion detection evolution offline dataset developed by MIT Lincoln laboratory. NSL KDD dataset is large enough in concerns of both number of instances and number of features, and it provides interesting characteristics on the distribution of events and on the dependencies between features.

The dataset contains training data and test data where training data include seven weeks of network traffic in form of Transmission Control Protocol dump data consisting of approximately 100 bytes. The test data included 2 weeks of traffic, with approximately 2 million connection records. Each connection record contains 7 discrete and 34 continuous features for a complete of 41 features and each record is labeled as either attack or normal. There are 21 classes of attack that fall under 4 types of attacks categories:

Table 1. Attacks in NSL-KDD dataset and their categories

Attack Class	Attack Name
DoS	Neptune, Pod, Teardrop, Land, Back
R2L	Guess-password, Ftp-write, Imap, Phf, Multihop, Warezclient
U2R	Buffer-overflow, Load- module, Per, Rootkit, spy
Probing	Port-sweep, IP-sweep, Nmap, Satan

Denial of Service (DoS) Attack: DoS Denial of Service (DoS) attack results by preventing legitimate requests to a network resource by intense the information measure or by overloading computational resources.

Probing Attack: Probing is a category of attack in which an attacker scans a network to collect information of target system prior to initiating an attack.

User to Root (U2R) Attack: In this type of attack, an attacker initiates the access with normal user account on the system and then able to utilize the system vulnerabilities to achieve root access to the system.

Root to Local (R2L) Attack: In this attack class, an attacker who does not have any account on a remote machine. he sends packet to that machine over a network and utilize vulnerabilities to gain local access as a user of that machine.

IX. LITERATURE SURVEY

Yi Yang *et al.* [2016] is about IDS incorporate physical knowledge, logical behaviors and , specifications protocol to give the solutions which should be comprehensive and effective just to mitigate various cyber attacks. The proposed approach comprises access control detection, protocol white listing, model-based detection, and multi-parameter based detection. This SCADA of specific IDS is validated and implemented which is utilizing a inclusive and actual test-bed of cyber-physical and data from a real 500kV [5].

M. B. Shahbaz *et al.* [2016] tells about, addressing the issue of reduction in dimensionality by proposing an efficient algorithm of feature selection that considers the correlation between a subset of various features and the behavior class label. Here the two correlation metrics are Correlation based feature (CFS) and symmetrical uncertainty (SU) that are utilize to examine the level of dependency between features and labels of classes, and among different features. Results after experiments , for NSL-KDD dataset shows that the approach which are having less features, the existing schemes significantly outperforms in concern of the training period time, and time which is taken to build the model, while it increases the accuracy of the system and proposed selection approach efficiency is tested on various classifying algorithms and results based on comparison, shows that J48 classifier with the highest accuracy and precision values and lowest miss rate and false alarm rate [6].

Muhammad Aksam Iftikhar *et al.* [2016] they present a computer aided diagnostic system utilizing Haralick texture features and statistical moments of intensity histogram based features for prediction of colon cancer grades from biopsy images. Further, Synthetic Minority Oversampling Technique (SMOTE) is applied to create a balanced dataset in respect to

equal representation of all the classes. The discerning features are selected from the oversampled feature space by using minimum Redundancy Maximum Relevance (MRMR) selection methodology. RBF which is Radial basis function here kernel of support vector machine (SVM) is used for classification of samples with colon cancer of three grades. The texture and statistic features groped with data balancing and feature selection enables the SVM classifier to achieve good performance. The individual effect of SMOTE and MRMR on performance enhancement has been analyzed in detail. The proposed cancer grade prediction system gains better efficiency than existing techniques on a colon cancer grading dataset in concerns of various performance measures such as accuracy, sensitivity, specificity, ROC curves, and Kappa statistics [7].

Mohammad Farid Naufal *et al.* [2015] it needs a technique to solve data imbalance problem using oversampling method. In Synthetic Minority Oversampling Technique (SMOTE) is used as oversampling method. It is having advantages of FARM in learning on dataset which is having quantitative attribute and combined with the selection process of software complexity metrics using CFS and oversampling using SMOTE, this method is expected has a good performance than the previous methods [8].

Datta H. Deshmukh *et al.* [2015] outlines a system in which a set of data-preprocessing techniques Feature Selection and Descritization are used. In dataset the high data dimensionality to its large feature set poses a significant challenge. In Preprocessing the Feature selection algorithm is utilizes for selecting require features. These activities help to enhance the precision of the classifier. After that many classifiers are utilize such as Naive Bayes, Hidden Naive Bayes and NBTree. The advantage of Hidden Naive Bayes is a data-mining model that luxuriate the conditional Independence assumption of Naive Bayes Method. The next Classifier utilizes NBTree that combines a hybrid of decision tree classifiers and Naive Bayes classifiers which considerably improves the accuracy of classifier and decreases the Error rate of the classifier [9].

D. P. Gaikwad *et al.* [2015] they use ensemble machine learning methodology to build effective IDS. The Bagging ensemble method with REPTree as base class is utilized to implement IDS. The required features from NSL KDD dataset are selected to enhance the accuracy of classification and to reduce the false positive rate. The experimental result shows that the Bagging ensemble using the REPTree as a base class exhibits greatest classification accuracy. The advantage of bagging method is, it has been taken less time to build the model. The proposed ensemble method provides competitively low false positives in comparison to other machine learning methods [10].

Choudhury *et al.* [2015] describe many classification methods and machine learning methods to classify the network traffic (normal, anomaly). They compare these algorithms and they have put forth improved machine learning methods that are necessary for proper detection of network intrusion. They find Random Forest and Bayes Net are suitable classifier for the intrusion detection and boosting is the best algorithm [11].

X. PROPOSED METHODOLOGY

This paper proposes to study hybrid classification method for identifying network attacks. The single classifiers are not generating accurate result for every attack type in IDS. NSL KDD data set suffering from class imbalance problem that is overcome by applying SMOTE (Synthetic minority oversampling technique) on the minority class .

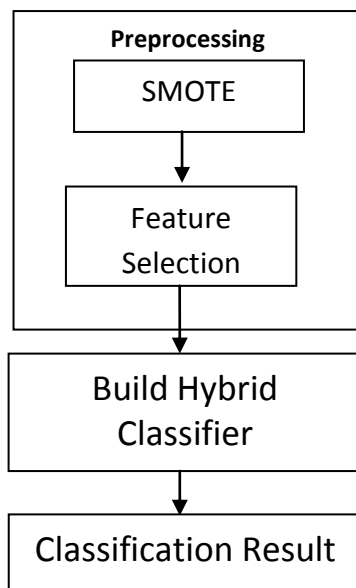


Figure 1. Implementation model

Table 2 shows the data distribution of NSL KDD in which some classes(minority classes) having very less number of records in comparison to other classes(majority classes). To overcome these difference SMOTE is applied on minority classes (spy, perl, phf, multihop, ftpwrite, loadmodule, rootkit, imap, land, warezmaster, buffer_overflow, guess_pwd, pod, warezclient, teardrop, back, nmap) that classes identified by K-Means clustering in minority cluster and oversampled these class data. K-means clustering result shows in table 3. SMOTE generated NSL KDD dataset shows in table 4 that assumed to be balance and utilize for feature selection. The Info gain feature selection method is applied on balanced dataset. The info gain feature selection is good for IDS. With the reduced feature list of dataset the hybrid

model is build for classification. The hybrid model contains The Random forest classifier and Random tree classifier which gives best classification results and the Byes Net classifier which is utilize for false positive reduction in intrusion detection system. The hybrid classification model generates more accurate results for U2R and R2L attacks in comparison to single classifier. Our implementation model consists of three stages shown in figure 1.

IX.1. Preprocessing:

Data preprocessing is technique that involves transforming raw data into relevant format. Dataset is usually imbalance, inconsistent, and/or lacking in bound behaviors or trends, and is probably going to contain several errors. Data preprocessing may be a manifest methodology of breakdown such problems. Data preprocessing produce raw data for further processing. In this first stage we generate relevant dataset from the complete large dataset by applying supervised instance filter and generate filtered dataset. This filtered dataset is utilize for processing and selects the required feature by applying filter method.

IX.1.1 SMOTE:

Synthetic minority oversampling technique (SMOTE) is a supervised filter method is utilize for oversample the minority class data and help to build balanced dataset. Dataset in which the no. of records is equally distributed is measure as balanced dataset. SMOTE generated balanced dataset gives to another preprocessing technique.

IX.1.2. Info Gain Feature Selection:

In information gain based feature selection the information gain or entropy is calculate for each attribute for the output variable. Information gain values vary from zero to one. Zero for those attributes that are having no information and one for the higher information attribute. Attributes that will have a higher information gain value can be selected; whereas those will not have much information can be removed. The info gain is a measure of reduction in entropy of class variable after the value for feature is observed.

IX.2. Build Hybrid Classifier:

Hybrid classifier can be presented as a combination of different classifiers. The classification ability of data-mining algorithm are different, this why combining them may increase the efficiency of the system in term of accuracy. This paper utilizing Vote ensemble method to build hybrid classifier in which we combine Random Forest, Random tree and Bayes Net classifiers to predict the possible intrusion. Vote is effective ensemble algorithm used for classification. In voting the two or more sub-models are created and each sub-model predicts and combines predictions in some way. Each sub model allows voting on the prediction. The random forest and random tree are gives best classification accuracy

on NSL KDD dataset for intrusion detection and bayes net is utilize for false positive reduction in IDS.

IX.3. Classification Result:

At last stage the hybrid classifier generate result for each individual class that present in the dataset.

XI. EXPERIMENTAL SETUP AND RESULT

The WEKA machine learning tool is utilizing for implementing the hybrid model. In which vote ensemble classifier is utilize for building the hybrid classifier.

On SMOTE generated balanced NSL KDD dataset applying the info gain feature selection method from The 41 features the relevant 22 features are selected in keeping with their info gain value: 3, 4, 5, 6, 10, 12, 23, 24, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, and 41. The evaluation of performance of existing approach and proposed approach is depicted in table 5. That illustrates when only RF is applied the accuracy is 99.8682% and FPR is 0.002. Proposed hybrid classifier generate 99.9894% accuracy and FPR reduce to 0.00 and other performance measures Precision, Recall, F-measure also have been improved for the U2R and R2L attacks.

Table 2. NSL-KDD Training Data Distribution

Attack Type	No. of records
Normal	67343
Neptune	41214
Satan	3633
Ipsweep	3599

Portsweep	2931
Smurf	2646
Nmap	1493
Back	956
Teardrop	892
Warezcilent	890
Pod	201
Guess_pwd	53
Buffer Overflow	30
Wazermaster	20
Land	18
Imap	11
Rootkit	10
Loadmodule	9
Ftpwrite	8
Multihop	7
Phf	4
Perl	3
Spy	2
Total	125973

Table 3. K-MEANS clustering generated result

Majority Class	Middle Class	Minority Class
Normal, Neptune	Satan, Ipsweep, Portsweep, Smurf	Nmap, Back, Teardrop, Warezcilent, Pod, Guess_pwd, Buffer Overflow, Wazermaster, Land, Imap, Rootkit, Loadmodule, Ftpwrite, Multihop, Phf, Perl, Spy

Table 4. SMOTE generated NSL KDD data distribution

Majority Class		Middle Class		Minority Class	
Normal	67343	Satan	3633	Back	8604
Neptune	41214	Ipsweep	3599	Teardrop	8028
		Portsweep	2931	Warezcilent	8010
		Smurf	2646	Pod	1809
		Nmap	1493	Guess_pwd	477
				Buffer Overflow	270
				Wazermaster	180
				Land	163
				Imap	99
				Rootkit	90
				Loadmodule	81
				Ftpwrite	72
				Multihop	63
				Phf	36
				Perl	27
				Spy	18

Table 5. Comparative analysis of performance

Performance Parameters		Random Forest	Clustering + Smote + Random Forest	Clustering + Smote + Feature Selection + Hybrid Model
	Number Of Features	41	41	22
	Classification Accuracy	99.8682%	99.8986	99.9894%
TPR	Normal	1.000	1.000	1
	Buffer_overflow	0.800	0.993	1
	Ftp_write	0.500	0.903	1
	Guess_passwd	0.943	0.994	1
	Imap	0.727	0.990	1
	Loadmodule	0.333	0.938	1
	Multihop	0.286	0.937	1
	Perl	0.667	1.000	1
	Phf	0.750	0.972	1
	Rootkit	0.000	0.867	989
	Spy	0.000	0.944	1
	Warezmaster	0.750	0.983	1
FPR	Normal	0.002	0.001	0
Precision	Normal	0.999	0.999	1
	Buffer_overflow	0.828	0.982	1
	Ftp_write	1.000	1.000	1
	Guess_passwd	1.000	1.000	1
	Imap	1.000	1.000	1
	Loadmodule	1.000	1.000	1
	Multihop	0.667	0.967	1
	Perl	1.000	1.000	1
	Phf	1.000	1.000	1
	Rootkit	0.000	0.963	1
	Spy	0.000	1.000	1
	Warezmaster	0.882	0.989	1
Recall	Normal	1.000	1.000	1
	Buffer_overflow	0.800	0.993	1
	Ftp_write	0.500	0.903	1
	Guess_passwd	0.943	0.994	1
	Imap	0.727	0.990	1
	Loadmodule	0.333	0.938	1
	Multihop	0.286	0.937	1
	Perl	0.667	1.000	1
	Phf	0.750	0.972	1
	Rootkit	0.000	0.867	0.989
	Spy	0.000	0.944	1
	Warezmaster	0.750	0.983	1
F-Measure	Normal	0.999	0.999	1
	Buffer_overflow	0.814	0.987	1
	Ftp_write	0.667	0.949	1
	Guess_passwd	0.971	0.997	1
	Imap	0.842	0.995	1
	Loadmodule	0.500	0.938	1
	Multihop	0.400	0.937	1
	Perl	0.800	1.000	1
	Phf	0.857	0.986	1
	Rootkit	0.000	0.912	0.994
	Spy	0.000	0.971	1

XII. CONCLUSION

In this work, to improve the performance of the classifiers and indentifying the intrusions in dataset the hybrid classifier is utilized. The relevant features are selected based

on their information gain value. The effects of feature selection are elevated using the hybrid classification method based on random forest, random tree and bayes net. The results shows that the hybrid approach achieved better

classification accuracy and gives better Precision, Recall and F-measure for R2L and U2R attack types. Hybrid classification approach is best for Root-kit attack, where single random forest is not identifying this type of attacks.

REFERENCES

- [1] R. Venkatesan, "A Survey on Wireless Intrusion Detection using Data Mining Techniques", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume. 1, Issue. 1, 2014.
- [2] M. Gyanchandani, J.L.Rana, R.N.Yadav, "Taxonomy of Anomaly Based Intrusion Detection System: A Review", International Journal of Scientific and Research Publications, Volume 2, Issue 12, 2012.
- [3] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Volume 16, pp. 321-357, 2002.
- [4] S. L. Pundir, A. Amrit, "Feature Selection Using Random Forest In Intrusion Detection System", International Journal of Advances in Engineering & Technology, Volume 6, Issue 3, pp. 1319-1324, 2013.
- [5] Y. Yang, L. Gao, Y. B. Yuan, K. McLaughlin, S. Sezer, "Multidimensional Intrusion Detection System for IEC 61850 based SCADA Networks", IEEE ,Volume 32, Issue 2, pp. 1068-1078, 2016.
- [6] M. B. Shahbaz, X. Wang, A. Behnad and J. Samarabandu, "On Efficiency Enhancement of the Correlation-based Feature Selection for Intrusion Detection Systems", In the proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference(IEMCON), Canada, pp. 1-7, 2016.
- [7] M. A. Iftikhar, M. Hassan, H. Alquhayz, "A colon cancer grade prediction model using texture and statistical features SMOTE and MRMR", In the proceeding of IEEE 2016 19th International Multi-Topic Conference(INMIC), Pakistan , pp.1-7, 2016.
- [8] M. F. Naufal, S. Rochimah, "Software Complexity Metric-based Defect Classification Using FARM with Preprocessing Step CFS and SMOTE", International Conference on Information Technology Systems and Innovation (ICITSI), Bandung – Bali, pp.16-19, 2015
- [9] D. H. Deshmukh, T. Ghorpade, P. Padiya, "Improving Classification Using Preprocessing and Machine Learning Algorithms on NSL-KDD Dataset" International Conference on Communication, Information & Computing Technology (ICICT), India, pp. 245 , 2015
- [10] D.P.Gaikwad, R. C. Thool, "Intrusion Detection System Using Bagging Ensemble Method of Machine Learning", In the proceeding of IEEE International Conference on Computing Communication Control and Automation, India, 2015
- [11] S. Chaudhary, A. Bhowal, "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection system", In the proceeding of IEEE 2015International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp.89-95, 2015
- [12] A. Tesfahun, D. L. Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature

Reduction", In the proceeding of IEEE International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, pp. 127-132, 2013

- [13] D. K. Dagly, R.V. Gori, R.R. Kamath, D. Sharma, "Hybrid Intrusion Detection System Using K-Means Algorithm", International Journal of Computer Science and Engineering, Volume 4, Issue 3, pp. 82-85,2016.

Authors Profile

Rajul Richhariya pursued Bachelor of Engineering from RGPV University, Bhopal, India in 2013 and pursuing Master of Technology from M.I.T.S., Gwalior, India.



Amit Kumar Manjharwar pursued Master of Technology from S.A.T.I. Vidisha. He is Currently working as Assistant Professor In Department of Computer Science Engineering and Information Technology, M.I.T.S., Gwalior, India. His area of interest focuses on Data Mining.



Rajni Ranjan Singh Makhwana pursued Master of technology from MANIT Bhopal. He is currently working as Assistant Professor in Department of Computer ScienceEngineering and Information Technology, M.I.T.S., Gwalior, India. His area of interest focuses on Algorithm Design and Network Security.

