International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Multilevel Fraud Detection System using Voting Techniques

Anjali Agrawal, Mahesh Parmar

Abstract*: Fraud detection is an enduring topic that pose a threat to the information security systems. Data mining helps detect and rapidly identify fraud and take instant action to reduce losses. It analysis the various forms of attacks through classifications algorithms of data mining. In this dissertation the data set of NSL-KDD is examined in identifying anomalies in network traffic patterns and the importance of various classification techniques is studied. Firstly the data is classify in two classes Normal and Anomaly and after then the anomaly data categories in various Attack forms i.e Probe, U2R, R2L, DOS to detect the fraud. The analysis performed through feature selection and classification techniques present in WEKA tool of data mining. The features are extracted through feature selection (CfsSubsetEval, InfoGain, FilterSubsetEval and FilterAttributeEval) methods from 41 attributes to 11 attributes using CfsSubsetEval with best first Search and after extraction the features the classification algorithms(Naïve Bayes, RC, RT, RF and DT) applied for finding Accuracy. The best Result in terms of Accuracy is finding through the voting method using Random Committee and Random Forest is 99.94%.

***Keywords*: fraud detection, Data Mining, classification algorithms, NSL-KDD dataset, Anomaly.**

I. INTRODUCTION

Data mining (DM) is a advance techniques in the type of expository devices. DM is a multidisciplinary area that combines machine learning, measurement, innovation in databases, and simulated awareness. This process covers different stages: understanding business, understanding data, data planning, modeling, evaluation and deployment.[16]
Through the usage of advanced tool of data mining, millions of activities can be explored in spot patterns and discover different types of fraud activities. The NSL-KDD dataset was analyzed using different clustering algorithms in the WEKA DM tool [3]. Each attack can be viewed in the flow of data as anomaly network, so these types of new attack can be identified through anomaly detection methods. By using this type of methods, the data is categorised into Anomaly and Normal attack.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Anjali Agrawal*, CSE/IT Department, Madhav Institute Of Technology & Science, Gwalior, India. Email: anjaliagrawal311093@gmail.com
Mahesh Parmar, CSE/IT Department, Madhav Institute Of Technology & Science, Gwalior, India, Email: maheshparmarcse@gmail.com

Anomalies data in a data collection comprises only a tiny part of all data associated within the normal data, so data is categorised through implementing intelligent algorithms and build a model that can be utilize later for intrusion detection. Different papers choose different approaches to detect anomaly. Although each vulnerability is unnoticed, the data can be categorized as defective or normal, so either all attacks must be used individually, or the attack must be categorized into different forms of attack [1].
NSL-KDD data set includes four main attack categories, involving Denial-of-Service (DoS) attacks (denying lawful system requests),Probing attacks (information gathering attacks),Remote-to-Local (R2L) attacks (unauthorized local remote access), andUser-to-Root (U2R) attacks (unauthorized local super-user or root access). The data set of NSL-KDD is split into labeled and unlabeled records.
1. **DoS Attack**: Denial of Service (DoS) attack results by stopping lawful applications by bandwidth consumption or overloading computing resources to a network resource. Neptune, Smurf, Pod and Teardrop are examples of this.
2. **Probe Attack**: Probe Attack shows the network data that can be utilized by another attack. This is a kind of attack that collects data of targeted system before starting an attack and deliberately thwarts its security controls. Portsweep, IPsweep, Nmap, and Satan are examples of this.
3. **Remote-to-Local (R2L) Attack**: In this, an intruder who does not have a remote machine account sends a packet over a network to a computer and exploits certain vulnerabilities to obtain local access as a user of that machine, unlawful local access from a remote machine e.g. Guess password, Ftp-write, Imap and Phf.
4. **User-to-Root (U2R) Attack**: In this case, an attacker who starts accessing the system to a normal user account and can exploit the system's vulnerabilities to gain root access to the system is unauthorized access to the root privileges. E.g. Buffer overflow, load module for Perl and Spy.

The rest of the paper is arranged as follows: Section II Provides a short overview of various existing and emerging work-related techniques described under the title literature study. In Section III help to understand the data description of the dataset of NSL-KDD. In Section IV it discuss about methodology for dataset with different classification algorithms. Section V describe the Experiment result and report of analysis on different classification methods. Section VI ends with conclusion.

II. LITERATURE REVIEW

S. Revathi and Dr. A. Malathi [1] , it evaluated the NSL-KDD dataset which resolves issues of the KDD cup99 dataset. The analysis explains that the NSL-KDD dataset is perfect to compare various models of intrusion detection.

It can take time to evaluate intrusion patterns using all 41 attributes in the dataset and also reduces system performance degradation. CFS subsets are used to decrease the dynamics of the data. Experiments have been used by various classification techniques for datasets without and with drawbacks and it are clear that the higher test accuracy is found by random forest in comparison to all other algorithms.

Nerijus Paulauskas and Juozas Auskalnis [2] it discuss the different classifiers that were used and these classifiers were added to an ensemble model to enhance the outcomes of intrusion detection precision. It observed that the model's accuracy is improved when using C5.0, J48, PART classifiers and Naïve Bayes with attacks grouped. It received the highest increase in the Naïve Bayes classifier situation. Only the use of NSL-KDD train + data for testing & training is determined, while 99 percent was not recommended without pre-processing the accuracy. In this, it is possible to test the model's actual performance to identify new types of attack. Pre-processing is used to train the data set; the same pre-processing should be carried with the numerical features of the training data set to test data. This is necessary so that numerical characteristics of test information are not acquired to train information and in this event the value may fail to evaluate the trained model. More precise outcomes are obtained using the proposed model of the classifier.

L.Dhanabal, Dr. S.P. Shantharajah [3], discuss the analysis on the NSL-KDD dataset suggest that this is one of the best information for simulating and testing IDS performance. The CFS technique for dimensionality reduction decreases the time of identification and improves the accuracy rate. This analyses, performed using NSL-KDD dataset statistics and tables, enables the researcher understand the dataset clearly. It also appears that most attacks are initiated using the TCP protocol's intrinsic deficiencies. An exploration on the possibility of using optimization techniques to create a model for intrusion detection with a better precision rate is suggested in the future.

Mubarek, A. M., & Adali, E[7]In this paper it has designed an intrusion detection and classification model through multilayer perceptions algorithm in comparison to naive and spear tree methods. The accuracy of this method exceeds the accuracy of complete data and is similar to the accuracy of other methods.

Teodora Sandra Buda1, Haytham Assem1 and Lei Xu1[17] The authors propose a strategic strategy to anticipate very timely defects or unusual trends within a set of data. Early detection of defects using the weighted anomaly window as a baseline to train the model, the ability to detect defects before they occur, prioritizes early detection. They learn different strategies for creating windows of fundamental truth. The results show that in the initial discovery score linked to each individual method, the ADE comprises an increase of at least 10% in all the information sets considered..

III. DATASET DESCRIPTION

Dataset description is the brief description of a dataset in a chronological form. It should be in a standardized format that enables facile exchange between data and its users [1]. NSL-KDD is a data set that is suggested to resolve some internal issues of the mentioned KDD'99 data set. NSL-KDD is a recent and standardized style of that has still suffer from some issues studied through McHugh and not be an idealized depiction of existing actual network, as there is insufficient public data sets for network based IDs, but It is still used and an effective standard to help researchers make comparisons between various intrusion detection methods from data is applied [2]. In addition, the number of instances in the test and train sets of NSL-KDD is acceptable. The advantage of NSL-KDD dataset is as this is cheaper to lead experiments on a full set without the requirement to randomly choose a fractional part. An improvement of the KDD'99 dataset to NSL-KDD dataset is that the NSL-KDD dataset has the advantage from the original KDD dataset is which as follows:

l. In this data redundancy or duplicate records is not include in the train set, then classifiers not be biased to similar records.

2. There is no duplication of record in the proposed test datasets, so performance of learners' is not biased in this ways it has better rates of detection on consecutive records.

3. The number of records are inversely proportional to the percentage of records, are selected by each hard level of groups in the original KDD data set.

4. The number of records in test and train sets is acceptable, which create this inexpensive to run experiments on a full dataset without the need to randomly select small parts [2].

Table I goes into detail shows the analysis of the NSL-KDD dataset in which number of individual records in both Normal and anamoly type attacks for both testing and training

Table I. Number of Instances in Normal and Anamoly Class

Dataset	All records	Normal	Anamoly
Train	125973	67343	58630
Test	22544	9711	12833

Feature analysis:

Feature engineering is using our knowledge of the problem to select features or create new features that allow machine learning algorithm to work more accurately. we create a new feature named attack type show in table, that describe the type of each given attack aiming to more clear observation and accurate detection to the result. This will be target value we are trying to predict with our model. We categorized into various types of attack, DOS, R2L, Probe and U2R from 39 various attack names. When training machine learning algorithms we always have to do two things with our data set. First shuffle the data so it is in random order, distribute the data in a training dataset and testing dataset [7].

Table II shows the various attacks in both testing and training dataset.

Table II. Attacks in Training and Testing Dataset

Attack Class	Attack Type
DOS	Udpstrom, Back, Processtable, Worm, Neptune, Land, Smurf, Pod, Teardrop, Apache2
Probe	Saint, Satan, Ipsweep, Mscan, Nmap, Portsweep,
R2L	Xlock, Xsnoop, Snmpguess, Phf, Imap, Multihop, warezclient, Warezmaster, Spy, Httptunnel, Snmpgetattack, Sendmail, Guess_Password, Named, Ftp_Write
U2R	Ps, Buffer_overflow, Perl, Xterm, Load_module, sqlattack, Rootkit,

Figure 1 shows the data flow diagram in which data is splitted in normal and anamoly data and then anamoly data is divided into following attacks.

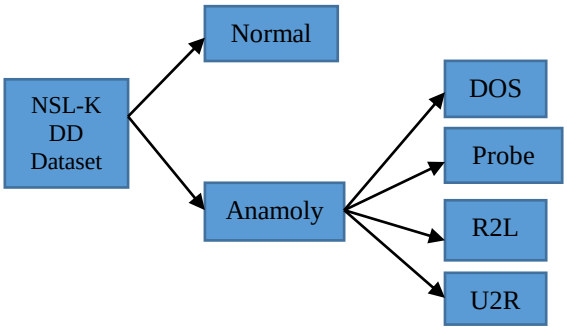

Fig 1. Data Flow diagram

TABLE III gives the information about the number of individual records in above forms of attacks for both the types of data set in training and testing.

Table III. Number of instances for different types of anamoly attacks

Dataset Type	DOS	Probe	R2L	U2R
Train	45927	11656	995	52
Test	7456	2421	2754	202

Figure 2 describe the table in this chart form to completely understand the divided data in attack types.

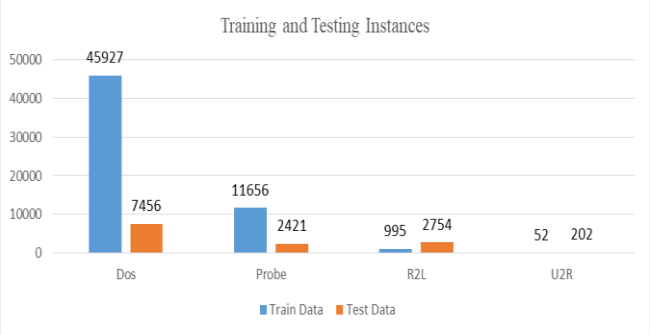

Fig 2 Number of Training and Testing Instances

IV. METHODOLOGY

To create the model of anomaly detection most of the paper used common feature extraction and classification techniques which are processed by following phases:

- Pre-processing of Dataset
- Feature selection.
- Create Model and evaluate it.
- Model Development.

In first step, the dataset is loaded and ready to select the features. If there is a need of standardized, normalized, redundant record of data can be removed.

Based on the methods used after the first step of pre-processing, using any machine learning algorithm features will be selected. Common approaches use wrapper or filter methods. Subset evaluator is used for wrapper method to generate all possible subsets by feature vector. Most Common subset evaluators are ConsistencySubsetEval and CfsSubsetEval. All features used in ranking order for the Filter method instead of showing the choosen characteristics. The feature last order having the least rank. The prediction model is built when the processed data is passed to classifiers. Greedy Stepwise, Best First search and Genetic Search algorithms are used to classify the wrapper method. The rank algorithm for the filter method is used to keep the characteristics from higher priority to least.

Compressive classification algorithms are numerous[4], Each with its own strengths and weaknesses. There is no single best-performing learning algorithm, but we selected five of them (innocent Buyers, Random Tree, Random Committee, Decision Table, Random Forest) in our analysis, and will try to find out which algorithm is best suited to the dataset [5]. Datasets are distinct in nature, so applying these algorithms through the Weka tool gives output outcomes that identify the correct algorithm for classifying information that performs better on various datasets. The Efficient classifiers are follows:

- Naïve Bayes.
- Random tree
- Random committee
- Decision Table
- Random Forest

Algorithms 1

Step 1: Load the NSL-KDD dataset
Step 2: Pre-process the dataset
Step 3: Apply the different feature Selection Method i.e CfsSubsetEval, InfoGain, FilterSubsetEval and FilterAttributeEval.
Step 4: Analysis the result of all feature selection methods and choose the best method i.e CfsSubset Eval.
Step 5: Again dataset loaded with best feature Selection method by CfsSubsetEval.
Step 6: Dataset is divided into two set i.e testing set and training set.
Step 7: The Classification Algorithms i.e Naïve Bayes, RC, RT, RF and DT are applied on training Dataset.
Step 8: Analysis the result of all classification algorithm applied and choose the best classification algorithms i.e RT

Step 9: Now the model built by Voting Method (RC, RT, RF) as Classifier

Step 10: Apply the test dataset on this classifier

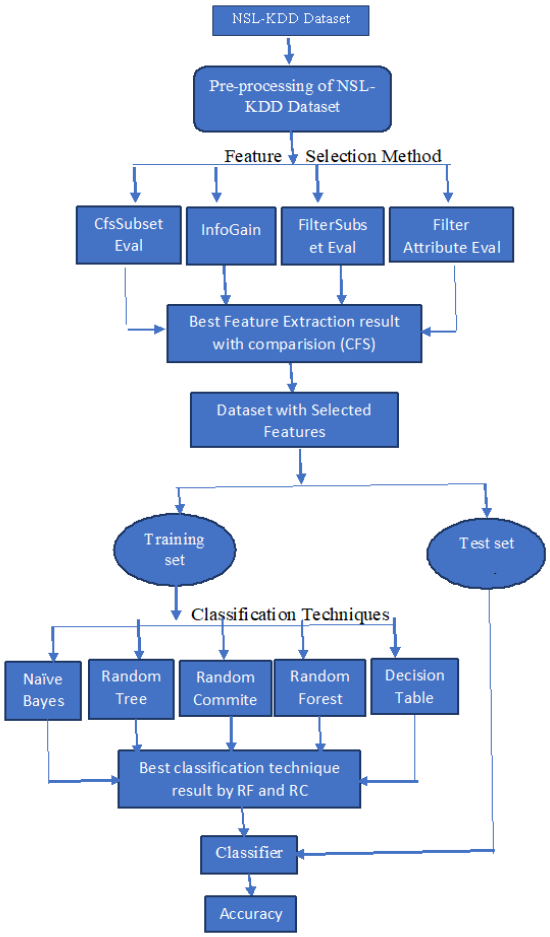

Fig. 3 Flow diagram of Methodology

Voting (ERCRTV) Method is used through the combination of Random forest and Random commitee for thr Accuracy of classification algorithm to perform model

Algorithm 2:

Step 1: while True do

Step 2: Load the Dataset of NSL-KDD

Step 3: Extract features using CFSSubsetEval Feature Selection Algorithm by Best-First Search method

Step 4: Combine Random Forest and Random Committee for the classification of samples using Voting Method.

Step 5: end while

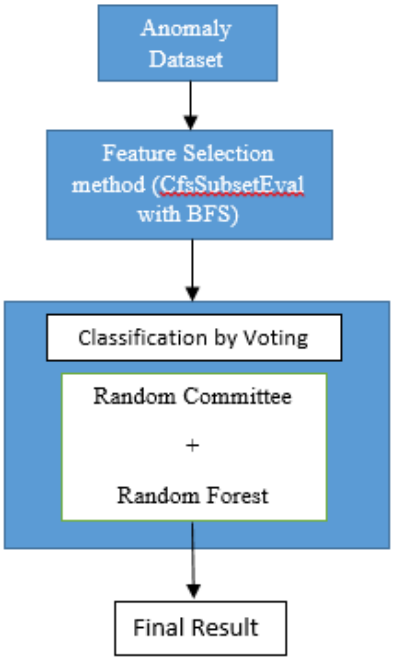

Fig 4. System Model for Propose Methodology

V. EXPERIMENTAL RESULT AND ANALYSIS

For our experiments, NSL-KDD datasets are chosen. It has both test sets and train cases, i.e. 22544 and 125973. Initially, The suggested form has been taught and tested using the train set and test set respectively. A tenfold cross-validation was conducted to evaluate performance metrics [4].

The information is either labeled as normal or one of 38 distinct kinds of attacks in the NSL-KDD dataset. It is possible to divide these 38 assaults into four groups: DOS, Probe, R2L, and U2R. The algorithm's efficacy is executed in the Weka tool [11]. It is a set of algorithms for information mining functions in machine learning. Developing new machine learning schemes is well suited[12]. WEKA comprises of four applications: Explorer, Experimenter, Flow of Knowledge, Simple Command Line Interface and Java Interface. The steps in the experiment are as follows

1. Select the dataset and pre-process it.
2. Run the algorithm for the classifier.
3. Compare the results of the classifier.

The first stage is pre-process discretization. The method of converting numerical attributes into nominal characteristics is discretization. The primary benefit is that some classifiers can take as input only nominal characteristics, not numerical attributes. Another benefit is that if the information is discreted before learning, some classifiers that can take numeric characteristics can attain greater precision. From the 41 attribute, we filtered to 12 function vectors using the CFS subset method to enable optimum selection for training and testing experiments from the entire dataset.

The table IV shows the comparison by applying all these four filter selection method.

Table IV. shows the feature selected by Filter methods

S.No	Feature Selection Algorithm	No. of Selected Features	Features Selected
1.	CfsSubsetEval + BFS	11	f3-f6, f12, f14, f29, f30, f37, f38
2.	InfoGain Attribute Eval + Ranker	24 (0.1)	f2-f6, f12, f23-f27, f29, f33, f35-f41
3.	Filter Subset Eval + BFS	15	f2-f6, f12, f14, f26, f29, f30, f35-f38
4.	Filter Attribute Eval + Ranker	20 (0.1)	f2-f6, f12, f23, ff25-27, f29, f30, f35-f41

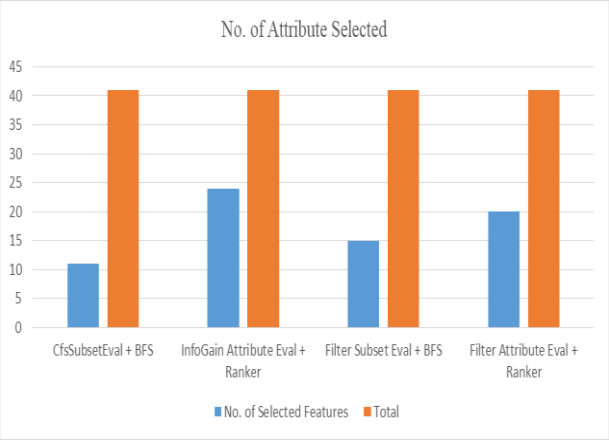

Fig. 5. Feature selected by Feature Selection Method

Table V shows the test accuracy on Anamoly Attack type class compared to 41 characteristics and the decreased set of characteristics using CFS subset method and shows the accuracy of the test obtained using the five algorithms. Here, by considering with and without feature reduction, the Random Forest and Random Commitee algorithm shows the highest accuracy compared to the rest of the algorithms.

TABLE V. Test Accuracy for Different classes of Attacks

Classification algorithm	Class name	Test accuracy with 12 features	Build time (in seconds)
Naïve Bayes	DOS	94.3	0.33
	Probe	91.2	0.15
	R2L	96.6	0.15
	U2R	93.8	0.19
Random Forest	DOS	98.6	3.02
	Probe	98.7	1.12
	R2L	97.8	1.33
	U2R	97.9	1.07
Random Tree	DOS	98.0	0.3
	Probe	97.7	0.15
	R2L	97.7	0.12
	U2R	98.8	0.12
Random committee	DOS	98.8	2.97
	Probe	98.7	0.89
	R2L	98.8	1.1
	U2R	98.9	1.29
Decision Table	DOS	98.7	1.18
	Probe	97.2	2.53
	R2L	98.6	0.61
	U2R	98.9	0.8

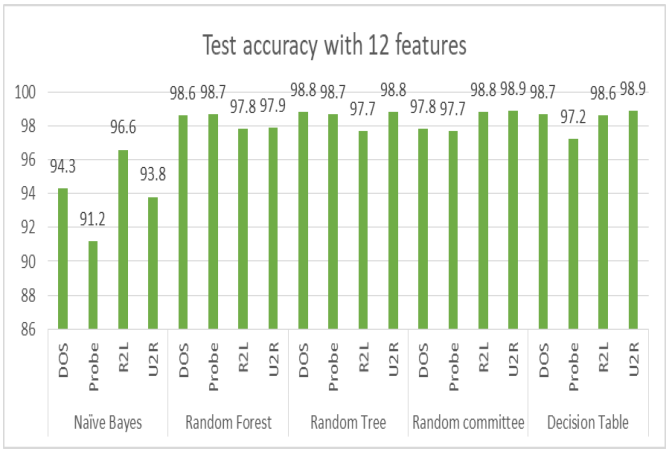

Fig 6. Accuracy of Different Classifiers

Table VI shows the train precision of the Anamoly Attack class using the combination of Random Forest, Random Tree and Random Commitee algorithm using Voting Methods indicates the greatest precision relative to the remainder of the algorithms.

TABLE VI. Accuracy of Attacks with Different Voting Approaches

Classification Algorithm	Class name	Test accuracy with 12 features	Build time (in seconds)
Random Forest + Random Tree	DOS	99.81	3.27
	Probe	99.73	1.14
	U2R	99.75	1.27
	R2L	99.90	0.99
Random Forest + Random Commitee	DOS	99.87	5.78
	Probe	99.77	1.84
	U2R	99.83	2.24
	R2L	99.94	2.26
Random Committee + Random Tree	DOS	99.81	3.94
	Probe	99.72	1.03
	U2R	99.75	1.17
	R2L	99.90	1.31

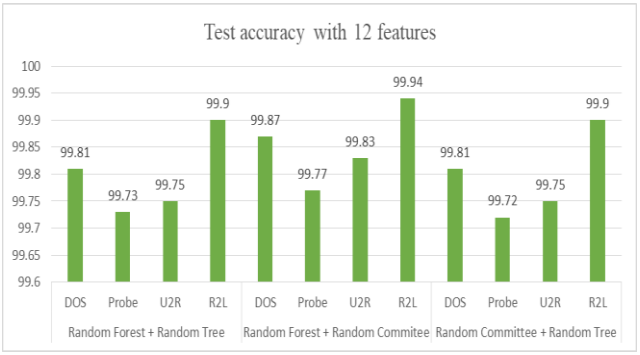

Fig 7. Accuracy of Different Classes of Attacks

Fig 7 shows the data that using Voting methods the Combination of Random Forest, Random Tree and Random Forest, but the highest accuracy is find through the Random Forest and Random Committee using Voting Methods.

VI. CONCLUSION

It present our concluding observations of the NSL-KDD dataset based on intrusion detection system. Attempts were made to reduce records and remove all unnecessary features, applying the feature extraction dataset as part of the pre-processing phase. Out of the 41 features in the NSL KDD, only 11 are selected using the CFS algorithm. The voting method is accessible in WEKA after removing less important attributes in this manner, it is used for a combination of the Random Forest and the technique of classification of the Random Committee. The suggested classification model also implemented a tenfold cross validation. With the use of voting method, the predictive precision of the proposed model is greater and the experimental results have shown that it consumes less construction time. The proposed model must also be tested on other datasets in order to reduce the total construction time required to complete the classification.

After reducing the attributes by using filter method the five classification algorithm is applied to check better accuracy and after then three algorithm is choosen. Secondly these all three algorithm is taken in voting method and applied the voting algorithm in Dataset to predict the accuracy of different forms of attacks of Anamoly.

The combination of Random forest and Random Committee classification techniques gives the best result as 99.4%

REFERENCES

1. SS. Revathi and Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" International Journal of Engineering Research & Technology (IJERT), IJERT ISSN: 2278-0181, Vol. 2 Issue 12, December – 2013.
2. NNerijus Paulauskas and Juozas Auskalnis, " Analysis of Data Pre-processing Influence on Intrusion Detection using NSL-KDD Dataset" 978-1-5386-3998-6/17/$31.00 ©2017 IEEE.
3. L.Dhanabal, Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015.
4. Niranjan A, Nutan D H, Nitish A, P Deepa Shenoy and Venugopal K R, "ERCRTV: Ensemble of Random Committee and Random Tree for Efficient Anomaly Classification using Voting" 978-1-5386-4273-3/18/$31.00 ©2018 IEEE.
5. Weka – Data Mining Machine Learning Software
6. Sandip r. shinde, poonam s. bhadikar, "A Genetic algorithm, information gain and artificial neural network based approach for hypertension diagnosis", international conference on Inventive Computation Technologies (ICICT), IEEE , 2016.
7. Mubarek, A.M., & Adali, E. (2017). Multilayer perceptron neural network technique for fraud detection. 2017 International Conference on Computer Science and Engineering (UBMK).doi:10.1109/ubmk.2017.8093417
8. Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB).doi:10.1109/aeeicb.2017.7972424
9. Mahesh Parmar: Comparative Analysis of Classification Techniques using WEKA on Different Datasets. International Journal of Latest Engineering and Management Research (IJLEMR). June 2018, Volume 03, Issue 06, PP. 01-05, ISSN: 2455-4847.
10. NSL-KDD, Dataset description, improvement to the KDD Dataset, [online] available http://www.unb.ca/cic/research/datasets/nsl.html
11. Anjali Agrawal, Mahesh Parmar: A REVIEW ON OPTIMIZATION TECHNIQUES USING DATA MINING. International Journal of Research and Analytical Reviews (IJRAR). February 2019, Volume 6, Issue 1, PP. 616-621, ISDN: E-ISSN 2348-1269, P- ISSN 2349-5138. (UGC Approved), (Impact Factor 5.75).
12. Kritika Yadav and Mahesh Parmar: Analysis of Mahatma Gandhi National Rural Employment Guarantee Act Using Data Mining Technique. International Journal of Computational Intelligence Research (IJCIR). 2017, Volume 13, Number 9 (2017), pp. 2221-2235, ISSN 0973-1873 (UGC Approved) (EBSCO database).
13. Kritika Yadav, Mahesh Parmar: Review Paper on Data Mining and its Techniques and Mahatma Gandhi National Rural Employment Guarantee Act. International Journal of Computer Science and Engineering (JCSE), April 2017, Volume-5, Issue-4,pp. 68-73, E-ISDN: 2347-2693. (UGC Approved)
14. Moksha Shridhar, Mahesh Parmar: Survey on Association Rule Mining and Its Application. International Journal of Computer Science and Engineering (JCSE), March 2017, Volume-5, Issue-3, pp. 129-135, E-ISDN: 2347-2693. (UGC Approved)
15. Monika Dandotiya, Mahesh Parmar: A Survey in Data Mining Prospective for handling Uncertainty and Vagueness. International Journal of Computer Sciences and Engineering (JCSE). 30 April 2019, Vol.-7, Issue-4, PP. 56-61,E-ISSN: 2347-2693, DOI:

https://doi.org/10.26438/ijcse/v7i4.5661. (UGC Approved),(Impact Factor 3.002).
16. N.Saravanan, Dr.V.Gayathri : CLASSIFICATION OF DENGUE DATASET USING J48 ALGORITHM AND ANT COLONY BASED AJ48 ALGORITHM. Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.
17. Teodora Sandra Buda1, Haytham Assem1 and Lei Xu1 "ADE: An Ensemble Approach for Early Anomaly Detection", Integrated Network and Service Management (IM), 2017 IFIP/IEEE Symposium, 2017.
18. P.Natesan1, P.Balasubramanie2, "Multi Stage Filter Using Enhanced Adaboost for Network Intrusion Detection" International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.3, May 2012

AUTHORS PROFILE

Anjali Agrawal, M.tech (IT) Student from MITS Gwalior
Area of Interest: Data Mining, Image Processing
E-mail: anjaliagrawal311093@gmail.com
I am pursuing M,tech in IT Stream from CSE&IT Department in MITS Gwalior. I received B.E. degree from SRCEM Gwalior. My area of current research includes Data Mining & Image Processing.

Prof. MAHESH PARMAR, Assistant Professor
B.E.(CSE), ME (Computer Engineering)
Area Of Interest: Data Mining, Image Processing.
E-Mail: maheshparmar@mitsgwalior.in.
Mr. Mahesh Parmar as an Assistant Professor in CSE&IT Department in MITS Gwalior and having 10 years of Academic and Professional experience. He received M.E. degree in Computer Engineering from SGSITS Indore. He has guided several students at Master and Under Graduate level. His areas of current research include Data mining and Image Processing. He has published more than 25 research papers in the journals and conferences of international repute. He has also published 02 book chapters. He is having the memberships of various Academic / Scientific societies including IETE, CSI, and IET etc.