Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms

Koushal Kumar
Assistant Professor
Department of Computer applications
Sikh National College Qadian (GSP)

Jaspreet Singh Batth
(Research Scholar)
Dept. of Computer Science and Engineering,
Guru Nanak Dev University

ABSTRACT
The task of developing Intrusion Detection System (IDS) crucially depends on the preprocessing along with selecting important data features of it. Another crucial factor is design of efficient learning algorithm that classify normal and anomalous patterns. The objective of this research work is to propose a new and better version of the Naive Bayes classifiers that improve**s** the accuracy of intrusion detection in IDS. The proposed classifier is also supposed to take less time as compared with the existing classifiers. To gain better accuracy and fast processing of network traffic**,** this study applied three standard methods of feature selection. This study tested the performance of the new proposed classifier algorithm with existing classifier**s,** namely Naïve bayes, J48 and REPTree thereby measuring different performance parameters using 10-fold cross validation. This study evaluates the performance of the new proposed classifier algorithm by using NSL-KDD data set. Empirical results of our study show that the proposed updated version of the Naive Bayes classifiers gives better results in terms of intrusion detection and false alarm rate.

Keywords
Machine learning; Intrusion Detection System (IDS); Naïve Bayes algorithm; Feature selection; NSL KDD data set

1. INTRODUCTION
The chances of data loss, hacking and intrusion have been increased with the growth and popularity of the Internet. Continuously growing Internet attacks pose severe challenges to develop a flexible and adaptive security oriented methods. An intrusion can be defined as a series of actions that compromises the integrity, confidentiality or availability of a computer resource [1, 2]. **Intrusion Detection System (IDS)** is one of the most important component being used to detect the Internet attacks that can be either host based or network based [3,4]. Intrusion detection is the process of monitoring and analyzing the activities occurring in a computer system or in a network in order to detect signs of security problems [5]. In literature, different techniques from various disciplines have been employed to develop efficient IDS but such techniques generally have some shortcomings. Traditional intrusion prevention techniques, such as firewalls, access control or encryption, password based security have been failed to fully protect networks and systems from increasing, sophisticated attacks and malwares [6]. The current trend in the research community to detect intrusion is based on machine learning. This technique has the capability of autonomous packet detections with high detection rates and low false positive rates while the system quickly adapts itself in dynamic environment. One of critical problem in network-based intrusion detection system is the extensive amount of data generated and collected from the network users. As the Internet users are growing exponentially, the data generated are also expanding day by day over computer networks and consequently the capability of IDS slows down [7]. The optimum feature set needs to be identified by extracting the unnecessary features which results in reduction in processing time and give higher detection accuracy in IDS [8, 9].

The present paper proposes an updated version of Naive Bayes (NB) classifiers that overcomes one of the drawback of existing Naïve Bayes algorithm and provides greater accuracy and preciseness in intrusion detection based scenario. For feature selection, this study has applied various feature selection techniques like Correlation-based Feature Selection, Information Gain feature evaluator, Gain Ratio attribute evaluation. The use of feature selection techniques removes the irrelevant or useless features that are not contributing much in intrusion detection. The proposed version of Naive Bayes classifier is tested on the NSL KDD dataset to detect attacks under the four main attack categories: Probe (information gathering), DoS (denial of service), U2R (user to root) and R2L (remote to local). The proposed version is also being compared with the existing classifiers. The rest of paper is organized as follows: Section 2 and 3 discuss about IDS with a brief introduction and feature selection techniques used in proposed work respectively. Section 4 gives a brief history of related work that has been done on NSL KDD data set in the field of Intrusion detection. Section 5 contains description of the new proposed Naïve base algorithm. Section 6 explains the employed data set and simulation environment. In section 7 discusses the results. Section 8 contains the conclusion and future work.

2. INTRUSION DETECTION SYSTEM AND MACHINE LEARNING
An Intrusion detection System (IDS) is defined as an effective security technology, which can detect, prevent and possibly react to computer related malicious activities [10, 11]. An ID monitors and analyzes statistics of networks activities for potential intrusion and security attacks [12]. IDS can be used to detect different types of malicious network communications and computer systems usage where as conventional techniques such as firewalls are easily vulnerable to attack and often prone to errors in case of wrong configuration or ambiguous security policies [13, 14, 15]. So to overcome the problems of traditional intrusion detection based approach Machine Learning (ML) based methods introduced. Machine learning is a field of Artificial Intelligence that is concerned with the design and development of algorithms that allow computers to learn with the help of example data [16, 17]. A major motive in machine learning research is to automatically learn to recognize complex patterns and rules so that it can make intelligent decisions based on the given data and past experiences. In

context of intrusion detection, a detection model learns from previously recorded attack patterns (called signatures) and detects similar ones in incoming traffic that have not been encountered previously [18,19,20]

3. FEATURE SELECTION METHODOLOGY

We can consider a number of factors affecting the success rate of IDS that is based upon machine learning classifiers on a given environment. One of those factors is representation and quality of the data that are we going to use for intrusion detection. Theoretically, having large amount of data with more attributes and features should result in more discriminating power and more accuracy. But practically, many machine learning algorithms have shown that this is not always true. Given a set of features, many learning algorithms produce a biased estimate of the probability of the class label [21]. If in the database there is too much irrelevant and redundant information present then learning during the training phase is more difficult. Redundant data directly lead to the problem of overfitting and the overall performance of the system will degrade. The Naive Bayes classifier can be affected by presence of redundant attributes due to its assumption that for a given class the attributes are independent. Decision tree algorithms such as C4.5 can overfit the training data, resulting in large tree size. Usually it has been seen that removing irrelevant and redundant information results in smaller tree production by C4.5 algorithm [22] [23]. So all the issues discussed above can be resolved using feature selection or attribute selection technique. Feature selection is used in intrusion detection to eliminate the redundant and irrelevant data. It refers to the process of selecting a subset of relevant features that fully describes the given problem with a minimum degradation of performance [24]. In attribute selection process the algorithm automatically searches for the best subset of attributes in your dataset. A subset of dataset is provided to algorithm for subset generation. Loop is implemented until it selects sufficient number of attributes from data set without influencing system performance and efficiency. A subset evolution function is used to for tracking this activity of algorithm. The whole process of feature selection is shown in Figure 1.

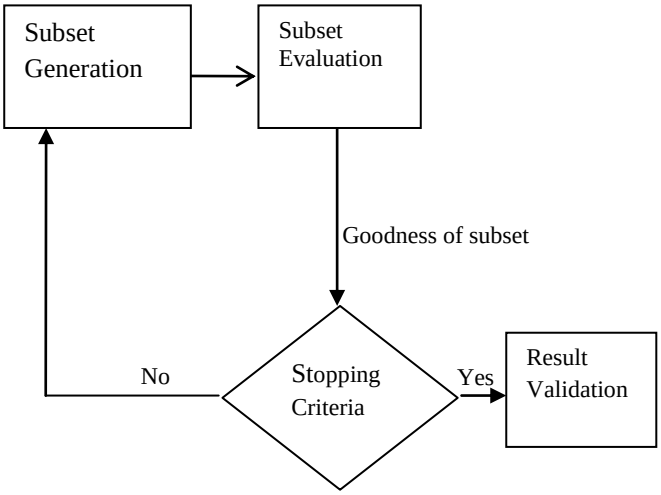

Figure 1: Procedure for Feature Selection

3.1 Attribute Selection

In this study three standard feature evaluator methods have been applied by us on our proposed new Naïve Bayes classifier algorithm. They are Correlation-based Feature Selection, Information Gain feature evaluator, Gain Ratio using best first search and rank based strategy.

3.1.2 Correlation-based Feature Selection (CFS):

It assesses the value of the group of attributes by concerning the individual predictive ability of each feature together with the possibility of repetition among the features. CFS attribute evaluator evaluates the features which are highly correlated with class, yet uncorrelated with each other [25]

$$R = \frac{\sum_{1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} \quad (1.1)$$

$$R = \frac{\sum_{1}^{N}(a_i b_i) - N\overline{A}\,\overline{B}}{N\sigma_A\sigma_B} \quad (1.2)$$

Here N is the number of tuples a_i and b_i is the respective values of A and B in the i^{th} tuple, A and B are the respective mean values of A and B, $\sigma_A and \sigma_B$ are the respective standard deviations of A and B. The value of R lies between -1 and 1.

3.1.3 Information Gain feature evaluator (IGF):

Information Gain Attribute Evaluation evaluates the worth of an attribute by measuring the information gain with respect to the class. Information gain is based on the concept of entropy which is widely used in the information theory domain. Given a collection of instances S, containing positive and negative examples of some target concept. The entropy of S relative to this Boolean classification is given by:

Entropy(S) = Info (G) = $-\sum_{i=1}^{m} p_i \log_2(p_i)$ (1.3)

Here G is calculated by calculating the probability of occurrence of class over total classes in data set, where p_i is the random probability that an arbitrary sample belongs to class C_i [26][27].

3.1.4 Gain ratio feature evaluator:

The gain ratio is an extension of information Gain discussed above. It attempts to overcome the Information Gain and prefers to select features that have a large number of values. So we can say Gain Ratio feature evaluator is more accurate under certain problems where data are well organized and there is no redundancy [28, 29]. Following equation calculates the value that represents the generated potential information by splitting the training data set.

Gain Ratio (A) = Gain (A)/Split Info (A) (1.4)

4. RELATED WORKS

Process of classification is widely discussed in the literature of intrusion detection process. Intrusion detection was manual task before 1985 with very poor chances of being able to detect intrusion. In 1980 the automatic intrusion detection concept began with Anderson's seminar paper [33]. He came up with a concept of a threat classification model. It employs a security monitoring surveillance system that is based on anomalies detection in user behavior. In 1987, Denning

proposed several models for IDS development based on statistics, Markov chains, time-series, etc. In Denning model, IDS identify the normal and malicious user's on the basis of their behavior like If a user behavior deviates sufficiently from the normal behavior is considered anomalous[34].The first IDS to achieve this in real-time was developed in the early 1990s. T. S. Chou et al. proposed a dynamic model of "Intrusion Detection System" based on one of specific Artificial intelligence approach like neural and fuzzy for intrusion detection. Chou et al. in their proposed model remove unwanted and ambiguous information from the network traffic [35]. Number of hybrid techniques has been used in machine learning field to overcome the problem of feature selection in intrusion detection. Hybrid approach based upon neural fuzzy or fuzzy genetic combine classification and clustering to enhance the performance of IDS [36] [37]. Al-Dabagh et al. show in their study that the accuracy and performance of IDS can be improved by selecting effective model of Artificial Neural Network (ANN) and its training parameters [38]. K Franke et al. proposes Correlation based Feature Selection (CFS) method which works automatically and effectively with nominal and continuous kind of attributes [39].Abraham A. et al. integrated Bayesian network and Classification & Regression Tee and proposed hybrid model for feature selection algorithms which gives better results in identifying unknown attacks [40]. Panda et al. proposed a hybrid intelligent approach using combination of data filtering along with a classifier to make intelligent decision that enhance the overall IDS performance [41]. Saurabh et al. showed the importance of features selection in building effective and efficient intrusion detection system. They proposed feature vitality based data reduction approach (FVBRM) to identify a reduced set of important input features using NSL-KDD dataset [42]. Alhaddad Mohammed J et al. carried out an experiment to study the applicability of different classification methods and the effect of using ensemble classifiers on the classification performance and accuracy [43]. Axellson proposes implication and base-rate fallacy for intrusion detection system that works on the principle of Bayesian rule of conditional probability [44]. E.N. Lutu proposed a Naive Bayes (NB) classifier for classification. Stream mining is the process of mining a continuous, ordered sequence of data items in real time. The performance of the Naive bayes classifier is improved by eliminating irrelevant features from the modeling process [45] [46] [47]. Xi-Zhao Wang et al [48] proposed a non-naive Bayesian classifier (NNBC) in which the independence assumption is removed and the marginal probability density function estimation is replaced by the joint probability density function estimation. Sanoop Mallissery et al [49] classify the NSL-KDD dataset with respect to their metric data by using the best six data mining classification algorithms namely J48, ID3 CART, Bayes Net, Naïve Bayes and SVM in order to find which algorithm will be able to offer more testing accuracy. S. Benferhat et al conducted an empirical investigation on the KDD Cup 99 data set, comparing the performance of NB and a Decision Tree (DT). The DT obtains a higher accuracy (92.28% compared with 91.47%), but NB obtains better detection rates on the three minor classes1, namely Probing, U2R and R2L intrusions [50]. You Chen et al. explore existing feature selection algorithms in intrusion detection systems group and compare different algorithms in three broad categories: filter, wrapper, and hybrid [51]. Stańczyk U proposed a new method of ranking for characteristics features using hybrid wrapper methods where ranking of variables is established based upon sequential backward search [52]. Datta H. Deshmukh et al. [53] developed a system which uses pre-processing methods like feature selection and descritization. Using feature selection algorithm required features are selected and due to descritization the data sets are discredited which is then applied to classifier algorithms like Naive Bayes, Hidden Naive Bayes.

which tries to distribute the classes evenly across the folds.

Step 2: Construct a primary structure of the Naïve Bayes classifiers with unique features, X={X1,X2,…..Xn}.

5. PROPOSED METHODOLOGY

Although Naïve Bayes algorithm gives satisfactory results with well organized data set but it always has some shortcomings like naive bayes has strong feature independence assumption i.e. it assumes that all features are independent of each other. From literature survey we come to know that a lot of effort has been done on improving Naïve Bayes classifier, following two approaches: selecting feature subset and relaxing independence assumptions are mostly used. In this study authors have proposed a new algorithm which works on both of above mentioned approaches. In the present work an updated version of the Naive Bayes classifier algorithm without assuming conditional independence of different attributes is proposed. The updated algorithm measures the relation between different attributes by applying the formula Corr(Xi, C) given in equation (1.5).In the next step we change the order like set X as a set X∗ in a descending order of |Corr(Xi, C)|.From the ordered features set X∗, an arc from the first features set is merged to the second features set. In the end, for all remaining features, we calculate the conditional probabilities of each feature with the help of previous features using the class values from the ordered set X∗. Maximum value of these conditional probabilities between all calculated features is used to distinguish the parent of each feature from its child. The correlation coefficients between two random variables X_i and X_j is defined as:

$$Corr(X_i, X_j) = \frac{N\sum_{i,j=1}^{N} X_i X_j - \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j}{\sqrt{(N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2 (N\sum_{j=1}^{N} x_j^2 - (\sum_{j=1}^{N} x_j)^2}} \quad (1.5)$$

Here N is the number of data points. This measure of correlation matrix has the characteristic of |Corr (Xi, Xj)| ≤ 1. When correlation value is close to 1, it shows the perfect linear correlation between variables Xi and Xj and Corr (Xi, Xj) = 0 means linear correlation is absent between features. The proposed algorithm works according to following steps.

Proposed Algorithm

Step1: Partioned the given data set into classes of approximate equal size. As the dataset contain classes of very unbalanced nature so proposed study has used k-fold cross validation.

Step3: Calculate the correlation and its types between each feature Xi, i = 1.....n of all classes using the correlation coefficient formula Corr(X_i Y_i) given in equation (1.8).

Step4. Change the order of elements of X as a set X^* = { $X^*_1, X^*_2, ..., X^*_n$ } in a decreasing order of |Corr(Xi,

C)|, i = 1, ..., n.

Step5. Add an arc from set X^*_1 to set X^*_2

Step6. For j = 5,6… ... n:

Step7 Find X∗ i that has the maximum value of $\sum_{k=1}^{N}$ | $P(X^*_{ki}, X^*_{kj}|C) - P(X^*_{ki}, X^*_{kj}|\overline{C})$ |,i<j Where $X^{*i} = (X^*_{1i}, X^*_{2i}, \ldots\ldots\ldots X^*_{Ni})^T$ here N is the number of elements in matrix and C = −C.

Step8 Join an arc from set X^*_i to set X^*_j.

Step9. Design new the conditional probability matrix deduce by the new structure.

Figure 2 shows the flow graph of the proposed algorithm where the NSL KDD dataset go through the phase of preprocessing and cross validation. Data is divided into training and testing sets and divided into K-sets. The proposed study uses k-fold cross validation method to distribute the classes evenly across the folds. Conditional and Posterior probability is calculated for each class of attack for improving its classification accuracy using proposed algorithm.

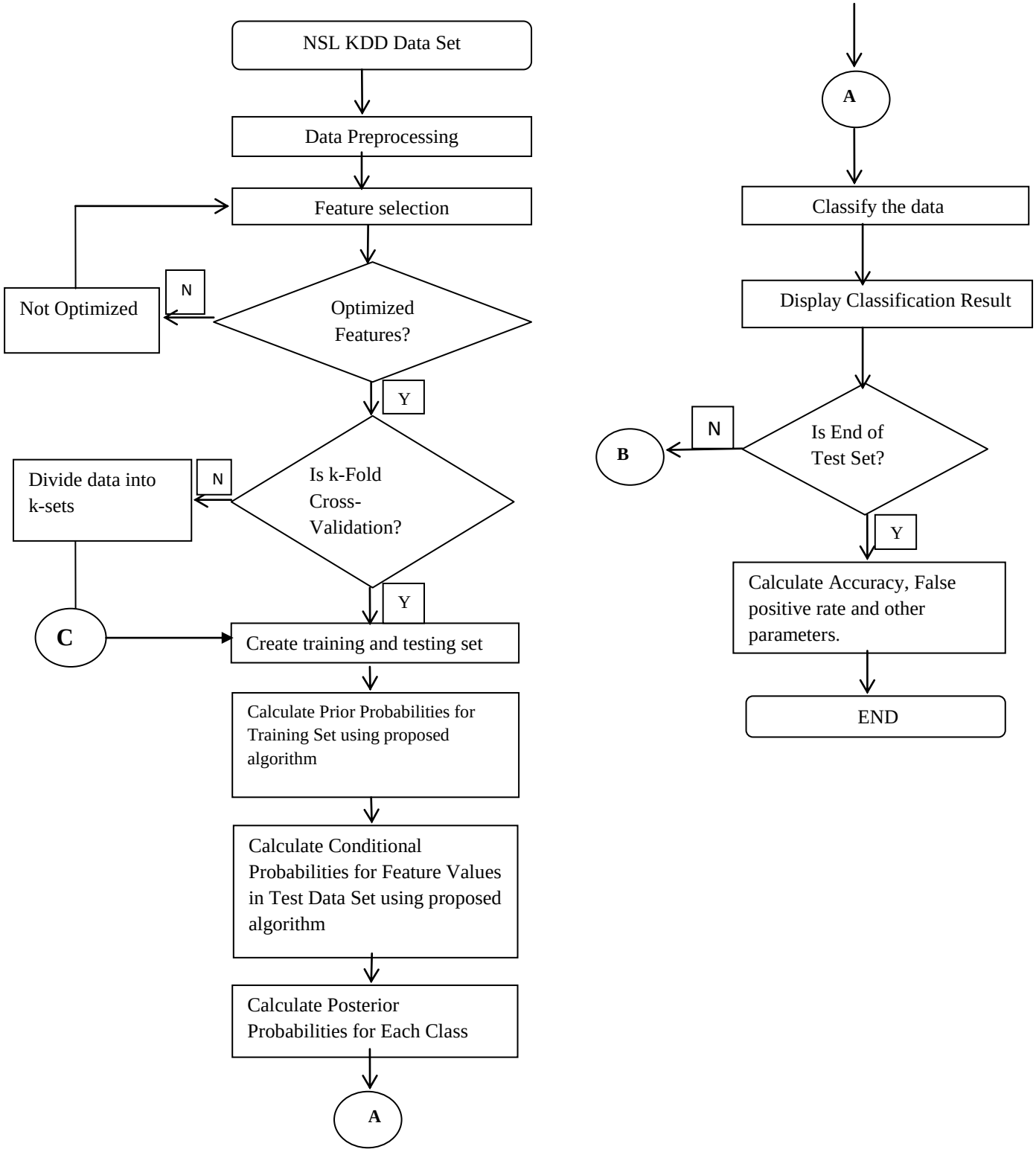

Figure 2: Flowchart of the Proposed Algorithm

6. DATA SET DESCRIPTION AND SIMULATION ENVIRONMENT

For experimental study we have used NSL-KDD data set which is an improved version of KDD data set and consists of selected records of the complete KDD data set. The training dataset used for experimental purposes has 21 different attacks out of total 37 present in the test dataset.NSL KDD data set is made up of 41 different attributes and five classes one of which is normal and other four are types of attack. The attack types are grouped into four categories as shown in **Table I and Table II** shows different instances of data set present in training and testing data set of NSL KDD data set.

Table I. Attacks and its types in NSL KDD Data Set

Attacks in Dataset	Attack Type (37)
Dos	Back, Land, Neptune, Pod, Smurf, Teardrop, Mailbomb, Processtable Udpstor m,Apache2,Worm
Probe	Satan, IPsweep, Nmap, Portsweep, Mscan,Sa int
R2L	Guess_password,Ftp_write,Imap,Phf,Multihop ,Warezmaster,Xlock,Xsnoop,Snmpgue ss,Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_ overflow, Loadmodule, Rootkit, Perl ,Sqlattack, Xterm, Ps

Table II. Instances of data set

Class Type	Instances in Training Data set	Instances in Testing Data set
Normal	67343	9711
Dos	45927	7456
Probe	11656	2421
U2R	52	200
R2R	995	2756

Simulation Environment: Waikato Environment for Knowledge Analysis (WEKA) is a popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License. It contains a collection of machine learning algorithms for classification. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. Weka supports several standard data mining tasks, more specifically data preprocessing, clustering classification, regression, visualization, and feature selection. The second tool used for implementation purpose is Netbeans which is an integrated developing environment (IDE) written in the Java programming language. The Netbeans Platform allows applications to be developed from a set of modular software components called modules. Applications based on the Netbeans Platform, including the Netbeans integrated development environment (IDE), can be extended by third party developers. The proposed algorithm is imported in Netbeans using weka API. Netbeans 6.0 and Weka 3.6 are used for implementation, respectively. Following Steps are used to extract the source code from weka tool.

1. Create a directory for the source code, e.g., the following: /tmp/weka
2. Extract the source code from the weka-src.jar with any archive manager that can handle the ZIP file format into the directory you just created (don't forget to re-recreate the folder structure when extracting).
3. Create a lib directory, if necessary (on the same level as src)
4. Run the build.xml (above the src directory) from command-line for creating all the necessary directories: ant exejar..

7. EXPERIMENTAL RESULTS

Performance measures indices: The performance evaluation of the experiment is carried out in terms of Accuracy (A), Detection Rate (DR) and False positive Rate (FPR) the following equations:

$$\text{True positive rate (TPR)} = \frac{TP}{TP + FN} = \frac{\text{Correct Intrusions}}{Instrusion} \quad (1.6)$$

$$\text{False positive rate (FPR)} = \frac{FP}{TN + FP} = \frac{\text{Normal As Intrusions}}{\text{Normal}} \quad (1.7)$$

$$\text{True negative rate (TNR)} = \frac{TN}{TN + FP} = \frac{\text{Correct Normal}}{Normal} \quad (1.8)$$

$$\text{False negative rate (FNR)} = \frac{FN}{TP + FN} = \frac{\text{Intrusions As Normal}}{\text{Intrusions}} \quad (1.9)$$

Two additional performance metrics are also commonly used, referred to as accuracy and precision

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\text{Correct Classifictions}}{\text{All instances}} \quad (2.0)$$

$$\text{Precision or Detection rate} = \frac{TP}{TP + FP} = \frac{\text{Correct Intrusions}}{\text{Instances Classified As Intrusions}} \quad (2.1)$$

True positive (TP): Classifying an intrusion as an intrusion. The true positive rate is synonymous with detection rate.

False positive (FP): Incorrectly classifying normal data as an intrusion also known as a false alarm rate.

True negative (TN): Correctly classifying normal data as normal, it true negative rate is also referred to as specificity.

False negative (FN): Incorrectly classifying an intrusion as normal.

Kappa statistics- Used in assessing the degree to which two or more raters, examining the same data, agree when it comes to assigning the data to categories.

NSL KDD data set that we are using for intrusion detection in our study has 125973 instances and 42 attributes. The empirical result in Table III shows performances of different classifier without feature selection technique. Although our purposed algorithm is better than existing classifiers and its performance is directly proportional to number of relevant attributes present in data set. Table III shows results on the basis of binary class i.e. attacker or normal user.

Table III: Classifiers Performance without Feature Selection

Classifiers	Time Taken(see)	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy TP Rate	Accuracy FP Rate	Precision	F-measure
Naïve Bayes	7.79	85.24	14.61	0.456	0.134	0.831	0.812
J-48	49.73	89.73	11.33	0.967	0.003	0.877	0.917
REPTree	21.12	90.07	9.03	0.892	0.025	0.883	0.721
Proposed Algorithm	24.45	92.34	8.66	0.90	0.061	0.913	0.97

Table IV shows some of the others statistics of our experiment. It has been found that some algorithm give more accurate results for a specific type of attack so it motivate us to apply our proposed algorithm to a class of attack like DoS, Probe, U2R and R2L. As shown in Table IV Naïve bayes algorithm is better than others algorithms in finding mean absolute errors. Similarly J-48 is good in finding Relative absolute errors. On the other hands REPTree and proposed algorithm is better in finding kappa statistics.

Table IV: Types of Errors during Testing

Classifiers	Kappa Statistics	Mean Absolute Error	Root-Mean Square Error	Relative Absolute Error
Naïve bayes	0.86	0.965	0.305	0.34
J-48	0.994	0.067	0.512	0.97
REPTree	0.992	0.025	0.038	0.506
Proposed Algorithm	0.921	0.021	0.312	0.312

To increase the accuracy and decrease the time of our intrusion detection system we have applied feature selection technique on 41 features using filtering method. A major challenge is to choose appropriate features selection methods that can precisely determine the relevance of features to the intrusion detection task and redundancy between features. So study has applied feature selection technique which is a process of selecting a subset of relevant features for use in model construction. Feature selection is itself useful, but it mostly acts as a filter, muting out features that aren't useful in addition to your existing features. We have applied three most commonly used feature selection algorithms namely ***Correlation-based with best first, Information Gain, Gain Ratio*** with ranker method and best first search strategy and find each algorithm have different number of attributes based on their evaluation criteria. Now we again compare the proposed algorithm with existing classifiers to identify and predict which classifiers can distinguish between alerts, attacks and normal data with maximum accuracy and can reduce false alarm rate as minimum as possible. Table V shows the number of features selected in each feature reduction method. Table V shows the number of features selected by each feature reduction method. Table VI shows results obtained after applying different classifier on different classes of attacks.

Table V: Depicts the number of features selected by each feature reduction method

Feature Selection Technique	No of Attribute Selected	Selected Attributes
CFS+ Best First	10	4,5,6,12,18,23,26,29,30,37,
Gain ratio + Ranker	18	3,4,5,6,9,11,12,17,22,25,26,29,30, 32,25,37, 38,39
Info Gain + Ranker	20	3,4,5,6,12,23, 24 ,25,26,29,30,31 ,32,33,34,35 , 36 ,37,38,39

Table VI: Classifiers Performance after Feature Selection

Feature Selection Algorithm	Classification Algorithm	Time taken(sec)	Correctly Classified	Incorrectly Classified	Accuracy TP	Accuracy FP	Precision	F-measure
CFS+Best First	Naïve bayes	5.22	93.5	6.95	0.937	0.134	0.935	0.916
	J-48	17.04	95.56	4.44	0.951	0.105	0.956	0.996
	REPTREE	7.5	96.40	3.60	0.966	0.121	0.96	0.976
	Proposed Algorithm	8.67	97.20	2.80	0.98	0.101	0.962	0.976
Info Gain + Ranker	Naïve bayes	2.25	94.145	5.85	0.975	0.097	0.92	0.947
	J-48	6.47	97.765	2.335	0.97	0.120	0.97	0.972
	REPTREE	4.24	96.12	3.881	0.96	0.046	0.96	0.996
	Proposed Algorithm	4.22	97.02	2.981	0.971	0.023	0.97	0.991
Gain ratio + Ranker	Naïve bayes	3.46	97.99	2.004	0.974	0.011	0.993	0.983
	J-48	8.46	98.178	1.822	0.974	0.007	0.995	0.985
	REPTREE	6.34	95.78	4.06	0.95	0.126	0.957	0.992
	Proposed Algorithm	5.25	98.94	1.62	0.986	0.002	0.98	0.99

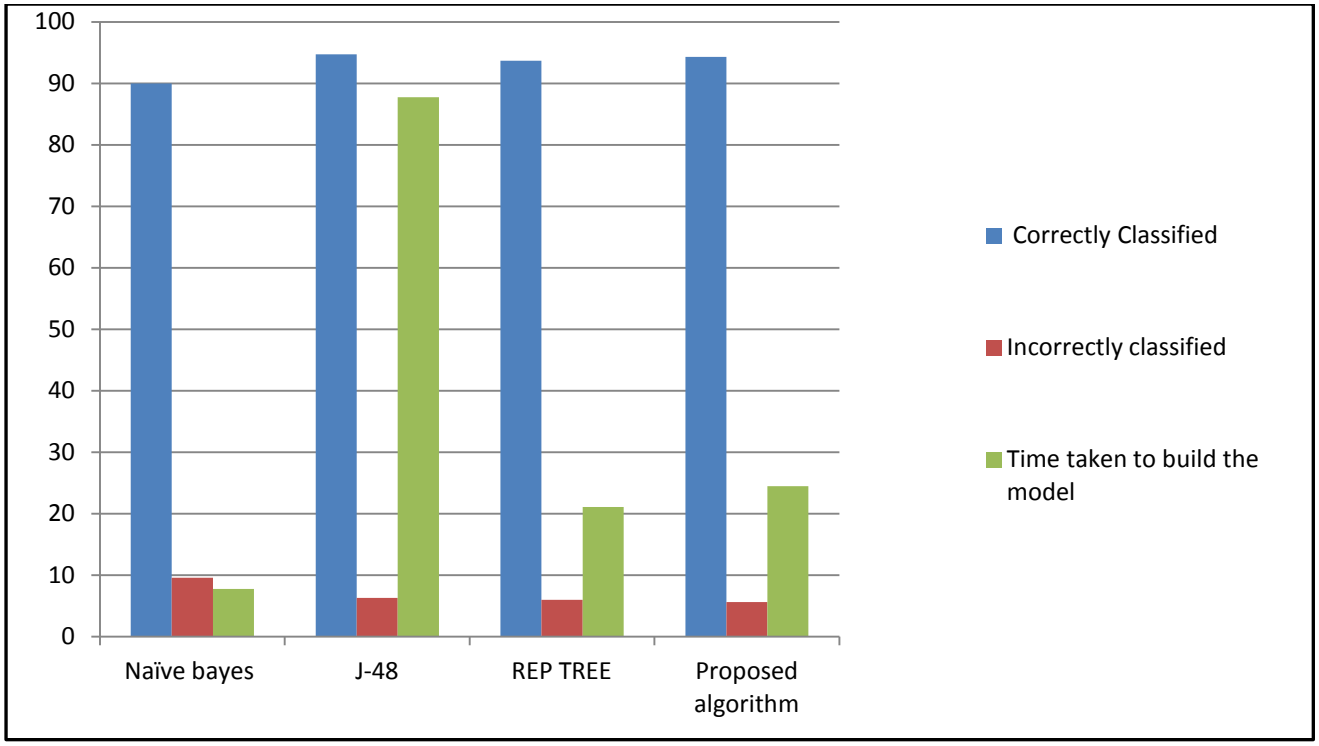

Fig 3: Comparisons of results before feature selection

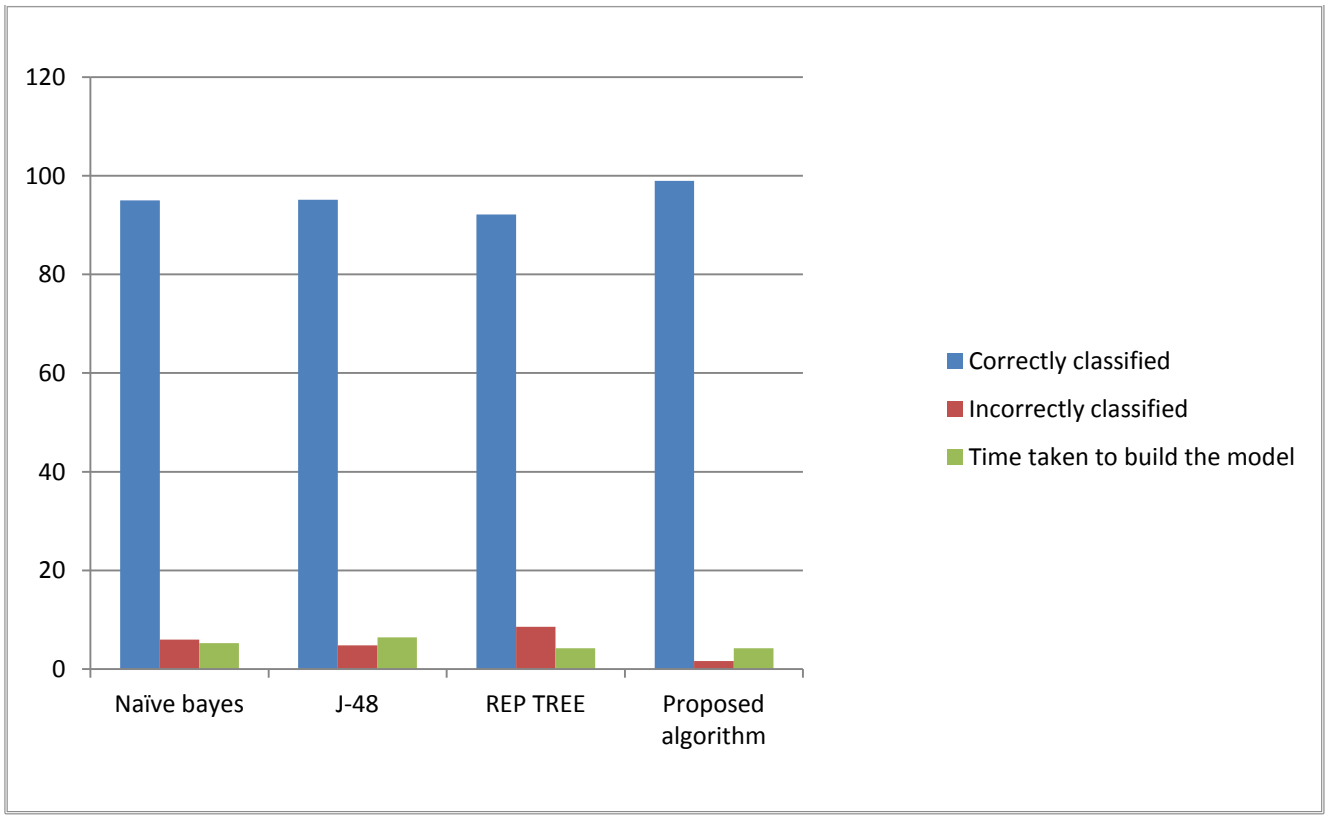

Fig 4: Comparisons of result after feature selection

Table VII: Test Accuracy for different classes of attacks

Classification Algorithm	Class Name	Test accuracy with 41 attributes in %	Test accuracy with 15 attributes in %
Naïve Bayes	Normal	95.127	96.864
	DOS	83.814	87.531
	Probe	89.517	90.245
	U2R	93.298	93.785
	R2L	86.782	87.748
J-48	Normal	97.622	97.895
	DOS	95.452	97.898
	Probe	97.865	98.165
	U2R	96.454	96.885
	R2L	94.231	96.434
REPTREE	Normal	96.982	97.564
	DOS	90.232	91.896
	Probe	85.452	85.762
	U2R	89.156	89.562
	R2L	90.512	92.762
Proposed algorithm	Normal	97.652	98.882
	DOS	96.452	97.521
	Probe	92.563	94.654
	U2R	96.788	96.882
	R2L	94.876	96.667

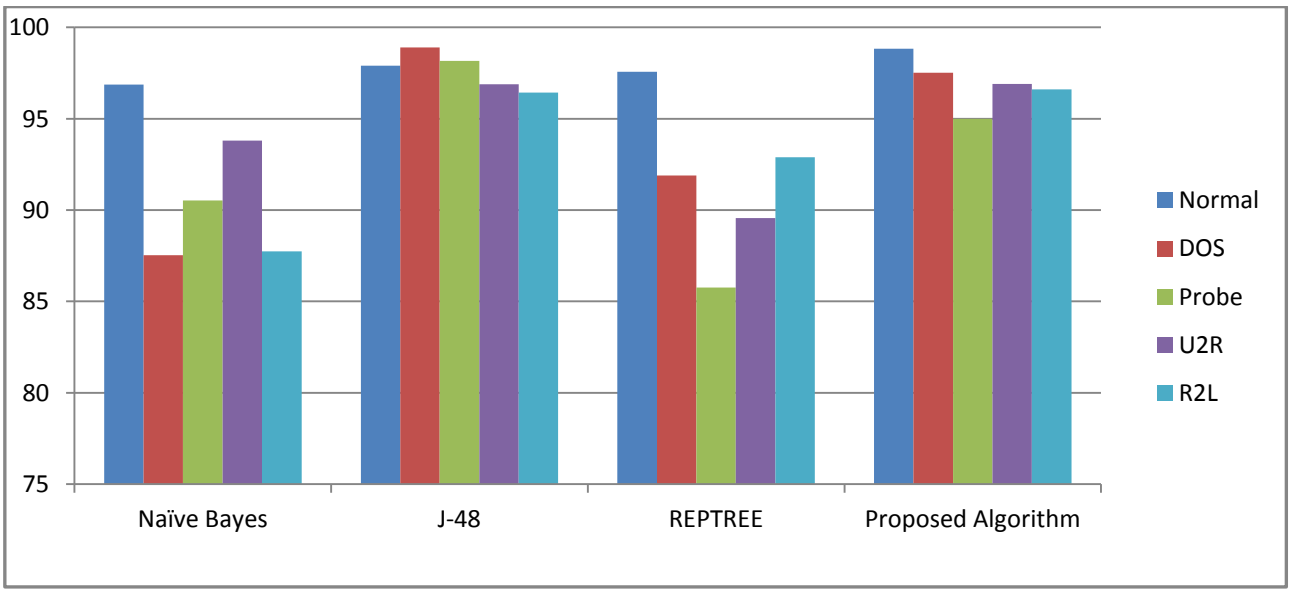

Fig 5: Comparisons of Test Accuracy for different classes of attack

7.1 Analysis of Results

Table VI has shown the results of all the feature selection algorithms used with proposed and existing machine learning classifiers. Result shows that proposed algorithm is better than existing Naïve bayes, J48 and REPTREE algorithms in terms of Root mean square error, Mean absolute error and Relative absolute error. So we can say less the error in detection more will be the accuracy of intrusion detection. **Figure 3** shows result in absence of feature selection approach which gives almost same accuracy in detection as other algorithms we compared in our research. **Figure 4** shows the results after feature selection approach and it is clear that in this case proposed algorithm gives more accurate results in intrusion detection and it also takes less time in identifying attacks. **Figure 5** shows results of best feature selection algorithms with three important parameters like correctly classified, incorrectly classified and time taken to build the model it also shows that Naïve Bayes classifiers gives maximum accuracy for U2R type of attack similarly J-48 and REPTREE gives maximum accuracy for Probe and R2L respectively. **Table VII** shows accuracy of classifiers on different classes of attacks using 41 and 15 attributes and it is clear from experimental result that Existing Naïve Bayes classify U2R attack more accurately then DOS, Probe and R2R attacks. Similarly J48 and REPTREE identify Probe and R2L attacks more accurately. Our proposed classifier is being able to detect every type of attack with more accuracy and preciseness as compare to others.

8. CONCLUSION AND FUTURE WORK

This study worked on issue related to Naïve Bayes machine learning classifier as it assume strong feature independence between attributes so proposed new algorithm which approximates the interactions between attributes by using conditional probabilities. The performance comparison amongst different classifiers with proposed classifier is made in order to understand their effectiveness in terms of various performance measures. From results, it is clear that every attributes in data set is not of equal importance, as we can ignore some attributes over others which does not involve much in intrusion detection. So this study has applied the feature selection techniques and found better results than before. Experimental result illustrates feature subset identified by Gain ratio + Ranker has improved our proposed Naïve Bayes classification. In future we will try to implement feature selection using soft computing techniques to identify intrusion in adaptive heterogeneous environment.

9. REFERENCES

[1] Chih-Fong Tsai a, Yu-Feng Hsu b, Chia-Ying Lin c, Wei-Yang Lin d "Intrusion detection by machine learning A review" Expert Systems with Applications Elsevier 2009.

[2] Tanya Garg and Surinder Singh Khurana IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur, India

[3] Jian Pei Shambhu J. Upadhyaya Faisal Farooq Venugopal Govindaraju. Proceedings of the 20thInternational Conference on Data Engineering published In IEEE 2004.

[4] Debar, H, Dacier, M., and Wespi, A, A Revised taxonomy for intrusion detection systems, Annales des Telecommunications Vol. 55, No.7–8, 361–378, 2000.

[5] Gulshan Kumar, Krishan Kumar & Monika Sachdeva (2010) "The use of artificial intelligence based techniques for intrusion detection: a review" Published online: 4 September 2010 © Springer Science+Business Media.

[6] Biesecker, Keith, Elizabeth Foreman, Kevin Jones and Barbara Staples (2008) "Intelligent Transportation Systems (ITS) Information Security Analysis." United States Department of Transportation Technical Report FHWA-JPO-98-009, 16 November 2008.

[7] Siva S. Sivatha Sindhu, Geetha , A. Kannan " Decision tree based light weight intrusion detection using a wrapper approach ".Expert Systems with Applications 39 (2012) 129–141 published in Elsevier

[8] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsin "A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment". Procedia Computer Science 3 (2011) 1237–1242

[9] F. Maggi, M. Matteucci and S. Zanero, "Reducing false positives in anomaly detectors through fuzzy alert aggregation". Information Fusion, 10, 300–311. 2009

[10] C-C. Lin and M-S. Wang, "Genetic-clustering algorithm for intrusion detection system. International Journal of Information and Computer Security", 2, 218–234. 2008

[11] Dr. Saurabh Mukherjee, Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Published by Elsevier 2012.

[12] O. Y. Al-Jarrah1, A. Siddiqui1, M. Elsalamouny, P. D. Yoo1, S. Muhaidat1, K. Kim "Machine- Learning-Based Feature Selection Techniques for Large- Scale Network Intrusion Detection" 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops.

[13] Heberlein, L. Todd, Dias, Gihan V, Levitt, Karl N, Mukherjee, Biswanath, Wood, Jeff, and Wolber, David, "A Network Security Monitor," 1990 Symposium on Research in Security and Privacy, Oakland, CA, pages 296-304

[14] Paxson, Vern, Bro, "A System for Detecting Network Intruders in Real-Time," Proceedings of The 7th USENIX Security Symposium, San Antonio TX, 1998.

[15] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey ," Intrusion Detection Using Data Mining Techniques" 978-1-4244-5651-2/10/$26.00 ©2010 IEEE

[16] PAT LANGLEY, STEPHANIE SAGE," Induction of Selective Bayesian Classifiers" Institute for the Study of Learning and Expertise 2451 High Street, Palo Alto, CA 94301

[17] ENGEN, "Machine learning for network based intrusion detection," Doctoral dissertation, Bournemouth University, 2010.

[18] S. Zaman and F. Karray, "Features selection for intrusion detection systems based on support vector machines," in Consumer Communications and Networking Conference, CCNC 2009. 6th IEEE, pp. 1–8, 2009.

[19] V. BoloN-Canedo, N. SaNchez-Marono, and A. Alonso-

Betanzos, "Feature selection and classification in multiple class datasets: an application to kdd cup 99 dataset," Expert Syst. Appl., vol. 38, pp. 5947–5957, 2011.

[20] T. O. Ayodele, "Types of Machine Learning Algorithms," in *New Advances in Machine Learning*, Y. Zhang, Ed., In Tech, 2010.

[21] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning (ICML '97); 1997; Nashville, Tenn, USA. Morgan Kaufmann; pp. 412–420.

[22] Karan Bajaj, Amit Arora, "Dimension Reduction in Intrusion Detection Features Using Discriminative Machine Learning Approach" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1, July 2013

[23] I.H.Witten, E.Frank, M.A. Hall, "Data Mining Practical Machine Leanrning Tools & Techniques Third edition", Morgan kouffman 2011.

[24] Sumaiya Thaseen, Ch. Aswani Kumar "An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System" International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 2013

[25] Mark A. Hall, "Correlation-based Feature Selection for Machine Learning" This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at The University of Waikato. April 1999

[26] Huy Anh Nguyen and Deokjai Choi "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model" APNOMS 2008, LNCS 5297, pp. 399–408, © Springer-Verlag Berlin Heidelberg 2008

[27] Dr. Saurabh Mukherjee, Neelam Sharma "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Procedia Technology 4 (119 – 128). Published under Elsevier 2012.

[28] Yogendra Kumar Jain, Upendra "Intrusion Detection using Supervised Learning with Feature Set Reduction" International Journal of Computer Applications (0975 – 8887) Volume 33– No.6, November 2011.

[29] Gulshan Kumar and Krishan Kumar "Design of an Evolutionary Approach for Intrusion Detection" Hindawi Publishing Corporation The Scientific World Journal Volume 2013, Article ID 962185, 14 pages

[30] Jun Li1, Lixin Ding and Bo Li "A Novel Naive Bayes Classification Algorithm Based on Particle Swarm Optimization" The Open Automation and Control Systems Journal, 2014, 6, 747-753

[31] C. Kruegel, D.Mutz, W. Robertson, and F.Valeur, "Bayesian event classification for intrusion detection," in Proceedings of the 19th Annual Computer Security Applications Conference (ACSAC 2003), 2003.

[32] Amor, Nahla B, Benferhat, S, Elouedi, Z. Naive Bayes vs. decision trees in intrusion detection systems. In: Proceedings of the 2004 ACM symposium on Applied computing, Cyprus, 2004, pp. 420–424.

[33] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980

[34] Dorothy E. Denning, "An intrusion detection Model," IEEE Transaction on Software Engineering", SE-13(2), 1987, pp. 222-232.

[35] T. S. Chou, K. K. Yen, and J. Luo "Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms. International Journal of Computational Intelligence 2008.

[36] Dewan Md. Farid a, Li Zhang a, Chowdhury Mofizur Rahman b, M.A. Hossain a, Rebecca Strachan, "Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks" Expert Systems with Applications Elsevier 2014.

[37] Mrutyunjaya Panda, Ajith Abraham, Manas Ranjan Patra "A Hybrid Intelligent Approach for Network Intrusion Detection" International Conference on Communication Technology and System Design 2011 Published by Elsevier.

[38] Saman M. Abdulla, Najla B. Al-Dabagh, Omar Zakaria, Identify Features and Parameters to Devise an Accurate Intrusion Detection System Using Artificial Neural Network, World Academy of Science, Engineering and Technology 2010.

[39] H Nguyen, K Franke, S Petrovic "Improving Effectiveness of Intrusion Detection by Correlation Feature Selection" , 2010 International Conference on Availability, Reliability and Security, IEEE Pages-17-24

[40] A Abraham, S Chebrolu, J P. Thomas "Feature deduction and ensemble design of intrusion detection systems" Computers & Security,Volume 24, Issue 4, June 2005, Pages 295-307

[41] Panda, Mrutyunjaya, Ajith Abraham, and Manas Ranjan Patra, "A Hybrid Intelligent Approach for Network Intrusion Detection," International Conference on Communication Technology and System Design 2011,Procedia Engineering 30 (2012), 1-9.

[42] Saurabh, Mukherjee, and Neelam Sharma, "Intrusion detection using naive Bayes classifier with feature reduction," Procedia Technology 4 (2012), 119-128 doi: 10.1016/j.protcy.2012.05.017

[43] Alhaddad, Mohammed, Amir Ahmed, Sami M. Halawani "A study of the modified KDD 99 dataset by using classifier ensembles approach," IOSR Journal of Engineering, May. 2012, Vol. 2(5) pp: 961-965.

[44] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection" Proc. Of 6th. ACM conference on computer and communication security 1999.

[45] Patricia E.N. Lutu, "Fast Feature Selection for Naive Bayes Classification in Data Stream Mining," Proceedings of the World Congress on engineering, Vol III, WCE 2013.

[46] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification," 2007.

[47] S Chebrolu, A Abraham, J P. Thomas Feature deduction and ensemble design of intrusion detection systems, Computers & Security,Volume 24, Issue 4, June 2005, Pages 295-307

[48] N. Ben Amor, S. Benferhat and Z. Elouedi. Naive Bayes vs. Decision Trees in Intrusion Detection Systems. In SAC ‟ 04: Proceedings of the 2004 ACM symposium on Applied computing, pages 420-424, New York, NY, USA, 2004. ACM. ISBN 1-58113-812-1.

[49] Xi-Zhao Wang, Yu-Lin He, Debby D. Wang, "Non-Naive Bayesian Classifiers for Classification Problems with Continuous Attributes" IEEE TRANSACTIONS ON CYBERNETICS 2013

[50] Sanoop Mallissery, Sucheta Kolekar, Raghavendra Ganiga "Accuracy Analysis of Machine Learning Algorithms for Intrusion Detection System using NSL-KDD Dataset" Future Trends in Computing and Communication 2013.

[51] You Chen, Yang Li, Xue-Qi Cheng, and Li Guo, "Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System", © Springer-Verlag Berlin Heidelberg 2006.

[52] Stańczyk U. Ranking of characteristic features in combined wrapper approaches to selection published in "Neural Computing and Applications". 2015 Feb 1; 26(2):329–44.

[53] Datta H. Deshmukh, Tushar Ghorpade, Puja Padiya, Improving Classification using Preprocessing and Machine Learning Algorithms On NSL-KDD Dataset, 2015.