False alarm reduction for Network Intrusion Detection System by using Decision Tree classifier

Sarah Mohammed Shareef

Computer science department of university of technology. Baghdad-Iraq

sarahshareef84@gmail.com

Dr. Soukaena Hassan Hashim

Computer science department of university of technology. Baghdad-Iraq

soukaena.hassan@yahoo.com

Abstract

Nowadays, Network security is one of the challenging issues with the rapid growth in information technology, this subject leading people to become increasingly aware of the threats to personal privacy through computer crime. Therefore, there is important to create intrusion detection system to detect malicious activities and various attacks on the internet with elevated detection rate and minimal false positive alarm. This paper proposed Network Intrusion Detection system using Decision Tree algorithm. To detect and classify attacks into four categories (DOS, Probe, R2L, U2R). The KDDcup99 dataset has been used to evaluate the activity of proposition system. The experimental results showed that the proposed system provides better results with high detection rate in experiment 1 (99.95%), experiment 2 (97.8%) and low false alarm rate in experiment 1 (0.05%), experiment 2 (2.2%).

Keywords: NIDS, alarm reduction, Kddcup99 dataset, Decision Tree

تقليل الانذار الكاذب لنظام كشف التطفل الشبكي باستخدام مصنف شجرة القرار

سارة محمد شريف **سكينة حسن هاشم**

قسم علوم الحاسوب،الجامعة التكنولوجية. بغداد-العراق

الخلاصة

في الوقت الحاضر، مع النمو السريع في تكنولوجيا المعلومات أصبحت امنية الشبكات واحدة من القضايا الصعبة مما جعل المستخدمين بان يكونوا على وعي متزايد من التهديدات للخصوصية الشخصية من خلال جريمة الكومبيوتر. لذلك، هناك أهمية لخلق نظام كشف تطفل للكشف عن الأنشطة الخبيثة والهجمات المختلفة على شبكة الإنترنت مع ارتفاع معدل الكشف وانخفاض إنذار إيجابي كاذب. هذا البحث اقترح نظام كشف تطفل شبكي باستخدام خوارزمية شجرة القرار. للكشف وتصنيف الهجمات إلى أربع فئات (DOS،Probe، R2L، U2R). لتقييم أداء نظام الاقتراح، تم استخدام بيانات KDD cup 99. وأظهرت النتائج التجريبية أن النظام المقترح يوفر نتائج أفضل مع معدل كشف عال ومعدل إنذار كاذب منخفض.

الكلمات المفتاحية: نظام كشف التطفل الشبكي، تقليص الانذار، مجموعة بيانات kddcup99، شجرة القرار.

1. Introduction

Today it is so serious to supply elevated level security to preserve highly critical and specific information. Intrusion Detection System is a fundamental technology in Network Security. These days many researchers have concerned on intrusion detection system utilizing Data mining mechanisms as an artificial proficiency [1].

Intrusion Detection System (IDS): is an appliance or software which checks network or device liveliness about bad activities and generates reports to an administration Station [2]. The techniques of IDS are divided in two categories: first one is Anomaly established on intrusion detection system was an equipment which detecting device malicious based on the "normal user profile" for utilized as a baseline and classifying it like each normal and abnormal. Second one is Misuse established on intrusion detection system was known as signing up -based detection because alerts have been created based on definite attack signing up [3].

Feature selection is the most significant preprocessing of data mining manners that utilized to recognize the irrelevant and abundant information and removing them as much as possible. Features can be defined as discrete, continuous or nominal. In general, features were identified below [4]:

1- Relevant : it mentions to the features that one have effectiveness on the product and their function cannot be supposed by the remainder.

2- irrelevant : it mentions to the features are specified as those features not holding every effectiveness on the output, and that values are formed at random for every symbol.

3- Redundant : the redundancy is occurred whenever a feature can hold the function of else.

Classification is data mining mechanism which is token every case of a dataset under sight and it is a supervised machine learning technique so it can touch classified data. . A classification based intrusion detection system will assort the entire network passing into either normal or abnormal. There are different classification techniques for example decision tree [5].

A Decision Tree (DT) is defined as a predictive modeling technique from the subfield of machine learning within the large field of artificial intelligence. It uses divide and conquer method for splitting according to attribute values. One of the most different decision tree algorithms are described as ID3 [6].

The ID3 algorithm is the basic algorithm of decision tree induction, it produces decision tree by means of compulsion in detail from the top to the bottom. It is used to construct the classification rules in the form of decision tree. This is utilized Shannon's entropy (ent) like a standard of choosing the significant attribute [7] [8]:

$$\boldsymbol{Entropy(ent)\ (s) = \sum_{i=1}^{c} - \ pi\ Log_2\ pi} \qquad \textbf{(1)}$$

Where:

pi considered the rate of patterns belong to the i th kind.

Information gain is generally utilized to determine the property for each node of generated decision tree by selecting the best feature at every step of rising a decision tree (DT). Information gain is calculates anticipated reduction within entropy occasion by awareness the amount of feature Fj. is utilized:

$$\boldsymbol{info\ gain(S.Fj) = Entropy\ (s) - \sum_{Vi \in V_{Fj}} \frac{|S_{vi}|}{|S|} . Entropy(S_{vi})} \qquad (2)$$

Where:

V_{Fj} was represented of whole potential amounts of attribute (Fj), (S_{vi}) is a subset of (S) about that feature (Fj) has value(*Vi*).

2. Related work

A survey has been achieved latest papers which implement training and testing of the system based on decision tree.

Anuar N.B et al., 2008, Design an organization to concentrate on discovery including statistical test of both anomaly and normal traffic instituted on KDDCup99 dataset, again the design involve a hybrid statistical proposition utilizes decision tree of data mining classification, The proposal proves that the decision tree for designing intrusion detection system is more suitable and accuracy than rule-based classifier [9].

Mukund Y.R et al., 2016, proposes the present mechanism for intrusion detection system to inform afflicted way of employing the HDFS (Hadoop Distributed File System) of machine learning algorithms, so to minimize the rate of false alarm, they were used decision tree technique and augment it in the operation with the multi-device capacity of the HDFS, therefore this approach was reduced the time taken by the DFS and improved the accuracy of the IDS [10].

Elekar K. SH. et al., 2015, executes various classifiers like C4.5 decision tree, Random Forest and Hoeffding Tree of intrusion detection then match the outcome utilizing WEKA. Outcomes display that a Hoeffding Tree awards superior result between several classifiers for distinguishing anomalies by the test data [11].

Xiang ch. et al., 2008, suggested system Design (hybrid classifier of multi-level for intrusion detection system) using Bayesian clustering and decision tree. Detection rate can be increased by implementing a fresh multi-level intrusion hybrid classifier. A model with 4 stages of classification is utilized for the metis classifier. The first level classifies the test data at 3 divisions (Denial of service, Probe, and Others). User to Root attack and Remote to Local attack and the Normal dealings are labeled as others in this level. Second level divides others into anomaly and Normal classes, while third level dis connects the Attack class from level 2 at User to Root attack and Remote to Local. Furthermore the fourth level classifies the attacks at more particular attack kinds [12].

3. Dataset Description

The KDD cup 99 dataset has been the point attraction for many researchers for evaluates intrusion detection algorithms [13]. It was prepared by Stolfo et.al. (Salvatore J.S., 2000). This dataset was consisted Tcp connections; each connection has 41 features with labeled

determine the type of a connection either normal connection or type of attack connection. The artificial attacks breakdown in the following four divisions (see table 1) [14]:

A. Access or Denial of Service Attack (DoS): it mentions to an attack that an intruder creates some counting or memory resorts turn on to shaft legal demands, or dismisses legal users' incoming to computer.

B. User to Root Attack (U2R): mentions to the category of deed that the intruder set out with incoming to normal employer computation at the instrument (maybe obtained by sniffing passwords) and was capable deeds some sensibility to earn origin incoming to the instrument.

C. Remote to Local Attack (R2L): happens when an intruder which has the capacity for transmitting packets to the computer through network but who does not have counting on which computer deeds several sensibility to acquire native access like a user of that computer.

D. Probing Attack: was a trial to collect data on the network of devices for visible intent of compassing its security dominance. Table 1 displays the four divisions and their corresponding attacks on every category.

Table 1: attacks Description of KDDCup99 Datasets

(4main) Attack type	Description	Attack classes
DOS	Denial of Service attacks	Neptune, apache2, back ,udpstorm, teardrop.
Probe	observation and probing	Satan, ip-ssweep, saint, Mscan, port-sweep.
R2L	Forbidden access from remote instrument to local instrument	Named, Xlock, send-mail,warezclient.
U2R	Forbidden access to local user priviledges by a local unpiviledge user	perl, spy, worm, Xterm, Http-tunnel.

4. Design of the proposal process

The suggested system is Network intrusion Detection System model depend on ID3 classifier to detect four types of attacks that threated the machines security, and to reduce false alarm rates in IDS by using well-Known dataset KDD Cup 99 datasets. Fig. (1) Describe the general structure of the NIDS Model.

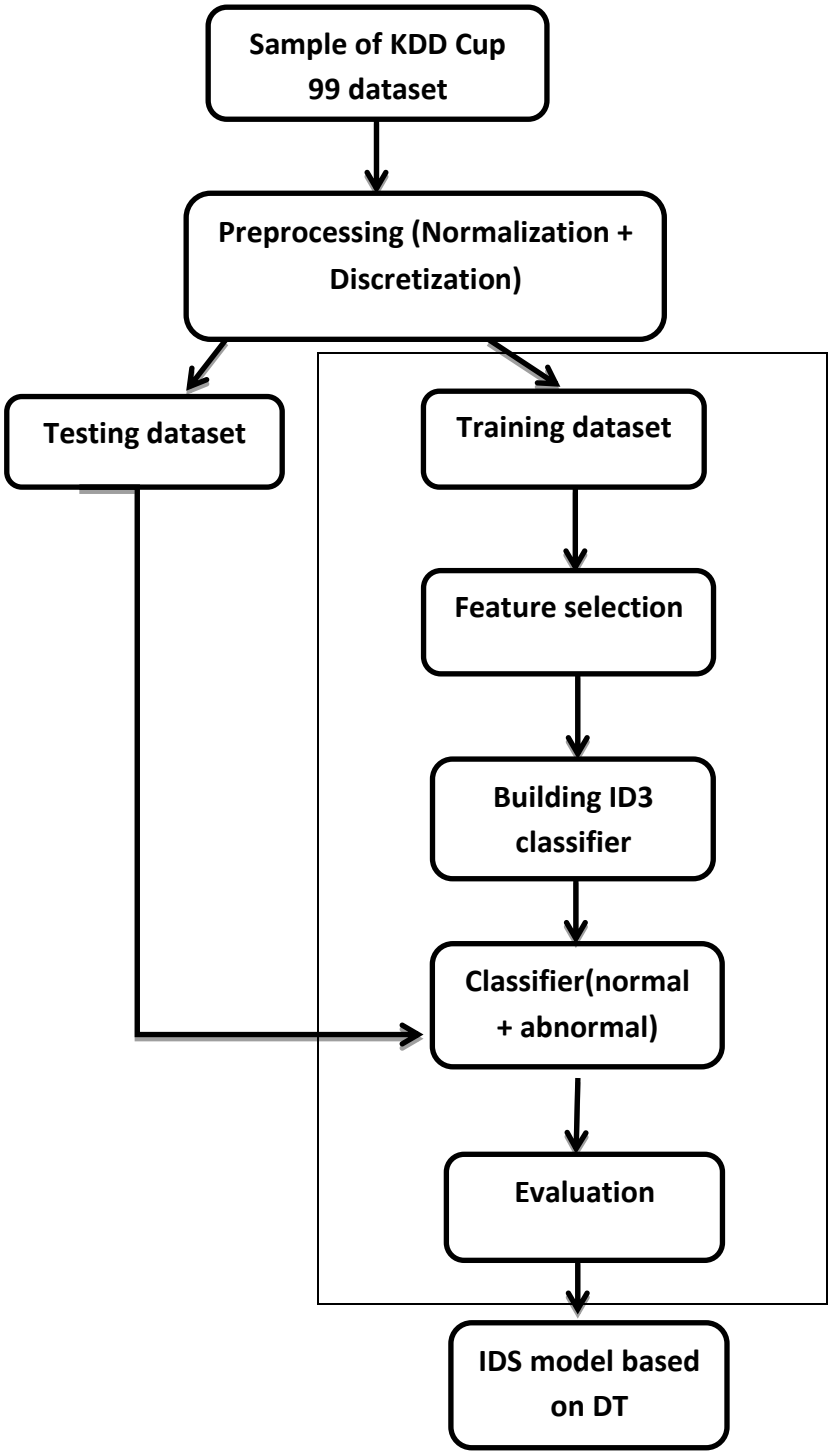

Figure 1. Block diagram for the proposed system

The proposed system illustrates in the following steps as shown in algorithm (1):

Algorithm (1) : The proposed system
Input: Training dataset **Output:** classify the type of attack samples from normal behavior
Begin Steps: 1- Preprocessing (Normalization and Discretization) of the selected subsets of samples. 2- Select subsets of samples for training and testing phases from Kdd99 dataset. 3- Feature selection method for reduction the redundant and irrelevant information. 4- Building ID3 classifier and training it by using the samples of training set. 5- Test the ID3 classifier using the testing samples. 6- Evaluate the performance of trained module. End.

4.1 Dataset preprocessing

Data preprocessing is a major and essential stage to obtain final datasets that can be considered beneficial for further data mining algorithms. In this proposal there are two types of data preprocessing are display as follows:

a) Normalize Dataset

First point after gaining the dataset from internet traffic, normalization process was applied to upgrade the action and efficiency of the system by scaling the accounts of feature during a small specified range [0 to 1].this proposal applied the normalization operation. Vision algorithm (2):

Algorithm (2) preprocessing of normalized dataset
Input: all Datasets of Continuous feature **Output:** values of dataset between 0 and 1
Steps: 1. For every attribute in dataset

compute the maximum value (max)
compute the minimum value (min)
for each value v in attribute

$$V' = \frac{V - min_A}{max_A - min_A}$$

End For
End For

b) Discretization dataset

Data Discretization was one of the basic reduction techniques adversary data preprocessing, In KDD cup 99 datasets contains continuous and discrete feature so it is serious to transform the continuous feature to discrete ones for guarantee the activity of the system. Discretization techniques are classified into: supervised and unsupervised discretization based on how it is performed, if the class information is utilized by the discretization operation then supervised is said. Otherwise it is unsupervised discretization.

4.2 Feature Selection Methods

To improve the effectiveness of the system reducing the consuming time must be used Feature selection technique for recognizing the irrelevant and redundant feature and removing them as much as possible. Feature selection techniques such as information gain, relief, gain ratio and the proposed system will be used entropy as feature selection.

4.3 The ID3 algorithm

ID3 algorithm is the basic algorithm of decision tree induction; it is used to construct the classification rules in the form of decision tree. Vision algorithm (3).

Algorithm(3) : ID3 classifier

Input: number of samples selected from KDD99 dataset (training dataset)
Output: set of classification rules
Steps:

1- For every class c in training sample
- Calculate p(c) from training sample
- Compute the entropy to all training dataset using Eq.1

end for.

2- For every attribute F in training sample using Eq.1
- Calculate the entropy

$$\boldsymbol{Entropy(s)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} - \boldsymbol{pi\, Log_2\, pi}$$

- Compute the Info gain using Eq.2

$$\boldsymbol{info\ gain(S,Fj)} = \boldsymbol{Entropy(s)} - \sum_{\boldsymbol{Vi \in V_{Fj}}} \frac{\boldsymbol{|S_{vi}|}}{\boldsymbol{|S|}} \boldsymbol{.Entropy(S_{vi})}$$

- Find the largest info gain
  Repeat until all entry values are empty.

3- If all classes are the same, then stop: decision tree has one node
Else goto on step 2

4- For every record in testing data:

1- Max=0, Class=""
2- For every Rule in training rule do steps 3,4
3- calculate Match that is a number of Rule conditions which is matched by record
4- If Match> Max
Then Max=Match, Class=class label of Rule
5- class of record is allocate by class label of Rule
End for
End for

End if
End for

4.4 Training and Testing of the proposed system:

In the learning stage the system used ID3 classifier on 4000 records for training operation by choose 1000 DOS, 700 probes, 200 R2L, 100 U2R and 2000 normal in KDDcup99 dataset.

In test stage 2000 samples are utilized to evaluate the work in KDDcup99 Dataset to establish the activity of the system, where the numbers of samples selected for each class demonstrate in (Table 2).

Table (2).dataset description

Number of dataset	Total number of records					
	records	normal	dos	probe	U2R	R2L
Train dataset	4000	2000 50%	1000 25%	700 17.5%	100 2.5%	200 5%
Test1 dataset	1500	337 22.4%	562 37.4%	254 17%	95 6.3%	252 16.8%
Test2 dataset	500	138 27.6%	113 22.6%	152 30.4%	19 3.8%	78 15.6%

5. Experiments and Performance Evaluation

The performance of the classifier used can be paralleled according to sure metrics such as accuracy, detection rate, and error rate, the confusion matrix is explained by four values which are TP, FN, FP and TN that shown in Table (3). The parameters are argued below.

True positive (TP): It mentions the number of attack which is detected as attack correctly.

True negative (TN): It mentions that number of normal which is predicted as normal correctly.

False negative (FN): It mentions the number of attack which is detected as normal correctly.

False positive (FP): It mentions that number of normal which is predicted as attack correctly.

$$(\mathrm{ACC}) = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

$$(\mathrm{DR}) = \frac{TP}{TP+FN} \qquad (4)$$

$$(\mathrm{FAR}) = \frac{FP}{TN+FP} \qquad (5)$$

Table (3): The confusion matrix

Actual class	Class Predicted	
	Normal class	Abnormal class
normal	(TN) True negative	(FP) False positive
abnormal	(FN) False negative	(TP) True positive

The KDDcup99 dataset used for training and testing ID3 algorithm, the suggested ID3 algorithm applied to classify dataset into five classes,(Normal,Dos,Probe,R2l,U2r). In the training phase (4000) records are elected randomly from whole dataset and used for training the algorithm .This subset of records contain normal and all other types of attack. To evaluate effectiveness of the proposed algorithm will conduct two experiments. In experimental 1 the trained model tested with (1500) subset of data of records contains both normal behavior and the four types of attacks. In experimental 2 subset consist of (500) record contains both normal and attack samples used to evaluate the proposed module. The subsets of data used in this work illustrated in table (2). Various performance measures used to evaluate the proposed module such as detection rate (DR), false alarm rate (FAR) and accuracy (ACC). The result obtain from testing phase show the high capability of proposed algorithm to distinguish normal activities from attack activities where the result from experiment 1 show effectiveness of the module to detect the attack behavior with detection rate reach to (99.95%), low false alarm rate reach to (0.05%) and accuracy of system is (98%). In experiment 1 the time for building model is (0.46) second and for testing model is (0.18) while In experiment 2 the time for building model is (0.35) second and for testing model is (0.13). The result from the two experiments shows in table (4) and figure 2.

Table (4) the experimental result

Performance measure	Exp1	Exp2
DR	99.95%	97.8%
FAR	0.05%	2.2%
Accuracy	98%	98%
True positive TP	99%	97%
False positive FP	0.4	0.5

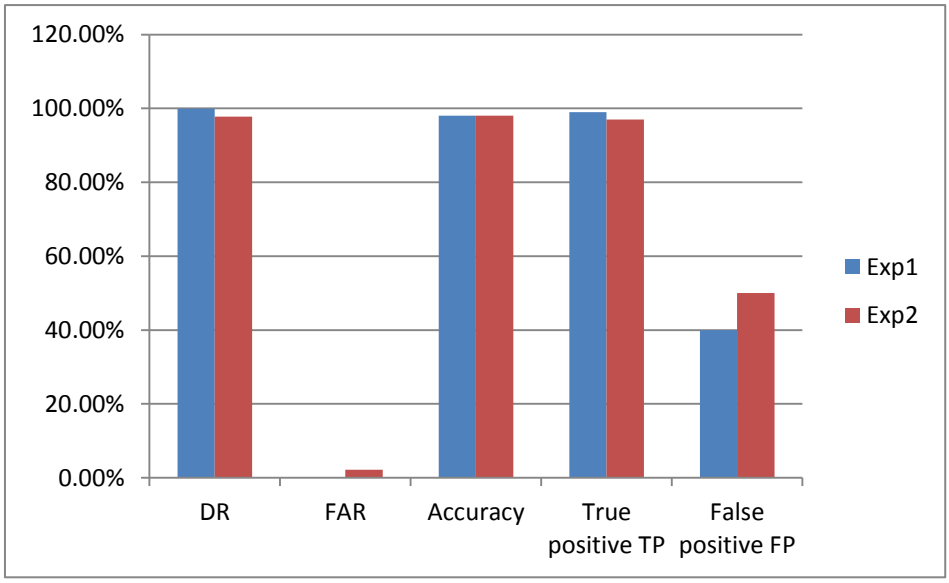

Figure .2: The results of performance measure

6. Conclusion

With the fast development in field of computer security, data mining mechanisms are also one of ways which security can be supplied to the network. In this paper network intrusion detection system based ID3 classifier algorithm proposed to detect more specific attack types from normal behavior. Data mining (DM) can assist to upgrade intrusion detection by processing every mentioned problem. In this work ID3 classifier algorithm used to classify the dataset into five classes. One for normal and another for attack behavior .The proposed module consist of two phase; Training phase where the module trained with dataset to be capable to distinguish normal behavior from attack event; Testing phase where new unseen samples utilized to evaluate the performance of module. The KDD99 dataset used to train and test the proposed module.experimental result display the effectiveness of system in detects attack and define normal behavior with 98% accuracy, high detection rate and low false alarm as shows in table (4).

References:

1- Khan J.A. and Jain N.,"A Survey on Intrusion Detection Systems and Classification Techniques", IJSRSET, vol 2, Issue 5, print ISSN: 2395-1990, online ISSN: 2394-4099,2016.

2- Kumar A., Maurya H.Ch. and Misra R.,"AResearch Paper on Hybrid Intrusion Detection System", International Journal of Engineering and advanced technology (IJEAT) ISSN: 2249-8958, vol-2, Issue-4, April 2013.

3- Gupta M., "Hybrid Intrusion Detection System:Technology and Development", International Journal of Computer applications(0975-8887), vol 115-No.9, Apr 2015.

4- Ladha L. and Deepa T., "FEATURE SELECTION METHODS AND ALGORITHMS", International journal on computer science and engineering (IJCSE) , ISSN:0975-3397, vol.3, No.5, May 2011.

5- Patil S.S., prof Kapgate D. and prof Prasad P.S., "A Review on Detection of Web Based Attacks Using Data Mining

Techniques", International Journal of Advanced Research in computer science and software engineering, ISSN:2277 128X, vol 3, Issue 12, December 2013.

6- Rajakshmi S.Ph.D and Shanthini J.S., "DATA MINING TECHNIQUES FOR EFFICIENT INTRUSION DETECTION SYSTEM: A SURVEY", International Journal on engineering technology and sciences- IJETS, ISSN(p):2349-3968, ISSN(0): 2349-3976, vol II, Issue XI, Nov 2015.

7- Hashem S.H. and Abdulmunem I.A., "A Model to Detect Denial of Service Attack Using Data Mining Classification Algorithms", thesis for master in computer science, june 2013.

8- Essa A.S., Orman Z. and Brifcani A.M.A., "A New Feature Selection Model based on ID3 and Bees Algorithm for Intrusion Detection System", Turkish Journal of Electrical engineering and computer sciences, Doi:10.3906/e1k-1302-53, 2015.

9- Anuar N.B., Sallehudin H., Gani A. and Zakari O., "Identifying False Alarm for Network Intrusion Detection System Using Hybrid Data Mining and Decision Tree", Malaysian journal of computer science, vol.21 (2), 2008.

10- Mukund Y.R., Nayak S.S and Chandrasekaran K.,"Improving False Alarm Rate in Intrusion Detection Systems using Hadoop", 2016 Inti. Conference on advance in Computing, communications and Informatics (ICACCI), India, Jaipur, sept.21-24, 2016.

11- Elekar K.Sh. and prof. Waghmare M.M.,"Comparison of Tree Base Data Mining Algorithms for Network Intrusion Detection", International Journal of Engineering, Education and technology (ARDIJEET), ISSN 2320-883X, vol 3, Issue 2, 01/04/2015.

12- Xiang Ch., yong P.Ch. and Meng L.S.,"Design of Multiple-level Hybrid Classifier for Intrusion Detection System Using Baysian Clustering and detection Trees", Elsevier, pattern recognition letters 29(2008) 918-924, doi:10.1016/j.patrec.2008.01.008.

13- Vijayarani S. and Maria S.S., "Introduction Detection System – Astudy", International Journal of Security, Privacy and Trust Management(IJSPTM) vol 4, No 1, Feb 2015.

14- Tavallaee M., Bagheri E., Lu W. and Ghorbani A.A, "A Detailed Analysis of the Kdd Cup 99 Data set", proceedings of the 2009 IEE symposium on computational intelligence in security and defense applications (CISDA 2009).