IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue5)

Available online at: www.Ijariit.com

An IDS by Correlation & KPCA with Neural Network Optimized By Genetic Algorithm

Harpreet Kaur
M. Tech. Scholar
Rayat Bahra University
Harpreetmultani6@gmail.com

Gaganpreet Kaur Bhalla
Assistant Professor, Rayat
Bahra University, Mohali
gaganb6@gmail.com

Abstract— An Intrusion Detection System is an application used for monitoring the network and protecting it from the intruder. Intrusion is a set of actions aimed to compromise these security goals. IDS have the computer security goals which are important for the data mining for extraction of data like confidentiality, integrity, and availability. This research study the performance measures of IDS is important for the security purposes. KDD 99 has 41 features. The IDS approached is used with the help of neural network technique of data mining and the Genetic algorithm is used as a classifier. The all features of KDD99 are used in this study. In this research, the feature is selected and extracted instead of using all features. The feature selection is done with the help of Correlation and feature extraction is done with the help of KPCA. The selection of features is according to the Eigen values of the features. The neural network is used for the change the weightage of the error. The neural network is basically runs many times and change weightage according to this.

Keywords— Data Mining, KDD Cup99, KPCA, Intrusion Detection System, Neural Network, Genetic Algorithm.

I. INTRODUCTION

An Introduction to Data Mining, Intrusion Detection System (IDS), and Neural Network (NN).

1.1. Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. The important data mining work is to extract the patterns of data from the huge amount of data that are present in any field like institutes, business etc. not from data itself. The data mining is the KDD process. For example, the data mining steps might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. There is no work for the individual working of the all processes of the KDD but all processes works together.

1.2. Intrusion Detection System

Internet and other web services are expanding very vastly in current years. Sometimes hackers or crackers misuse these services for the purposes or threat the data of that service or gaining some information for the security purposes. They harm the data of that web service and destroy it. So many techniques are established to protect the data like firewalls, antiviruses etc. but they could do it properly. So, the Intrusion detection system is used to protect the data from external forces that extract the data and harm it. Basically, we can say that IDS protects the computer security goals i.e. confidentiality, availability and integrity and IDS compromises according to these goals.

An ID basically is of two types: Anomaly and Misuse Intrusion detection system. In Misuse detection, attacks can be represented in the form of pattern or a signature in order to detect or prevent same attack in future. In anomaly detection category, deviation of normal usage behavior pattern is identified in order to correctly detect the intrusion.

How is data mining used in IDS? Pattern reorganization is the difficult task in the data mining. So, it can be handled by IDS and provide the classification is done by machine learning algorithms. Feature selection is the major problem in the leaning system. So, Ids is working with KDD to perform in the less set of features and improve the accuracy and performance of the system.

1.3. Neural Network

Neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

Neural network is made up of three layers i.e. input, hidden and output layers. A hidden layer is the processing layer. The hidden layer is in between the input and output layer.

Neural network on IDS is change the weightage of the input data according to the values.

This paper organized as follow: Section 2 present History, Section 3 also present KDD CUP 99 Dataset, Section 4 explain proposed approach, Section 5 consist experiments result, finally conclusion and future work is mentioned in Section 6.

II. LITERATURE REVIEW

The Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. A simulation study was performed on KDD-99 cup data set for intrusion detection. Various classifier algorithms are applied over the data set and a learning model is tested [3]. The identified set of all features of KDD cup is used with SVM and neural networks [4].

Chena *et al.* [33] have used a Flexible Neural Tree (FNT) model for the IDS. The FNT model reduces the number of features. Using 41 features, the best accuracy for the Denial of Service (DoS) and User Gain Root (U2R) is given by the FNT model. The decision tree classifier supply the best accuracy for normal and probe classes, which are an improved less than the FNT classifiers.

A Genetic Algorithm (GA) approach with an improved initial population and selection operator, to efficiently detect various types of network intrusions. In the testing phase the Network Security Laboratory-Knowledge Discovery and Data Mining (NSL-KDD99) benchmark dataset has been used to detect the misuse activities. By combining the IDS with Genetic algorithm increases the performance of the detection rate of the Network Intrusion Detection Model and reduces the false positive rate [24].

The feature selection is done with the help of information gain and genetic algorithm is used for extracted that. The accuracy is checks according to the attacks of the KDD99 cup with intrusion detection [29].

3. KDD CUP 99 Data Set

KDD Cup’99 dataset used for benchmarking intrusion detection problem is used in our experiment. These are generated by processing the tcpdump segment of DARPA 1998 evaluation data set. This data set consists of 41 feature and separate feature (42nd feature) that labels the connection as ‘normal’ or a type of attack. The data set contains a total of 23 attack, these are grouped into 4 major categories:

3.1. Denial-of-Service (DoS)

In Denial-of-service attack, the attacker has the goal of limiting or denying service provided to the user, computer or network. Attacker tries to prevent genuine users from using a service. It is usually done by making the resources either too busy or too full and overflow.

3.2. Probing or Surveillance

Probing or Surveillance attacks have the main aim of gaining knowledge of the existence or configuration of a computer system or the network. The attacker then tries to harm or retrieve information about resources of the victim network.

3.3. User-to-Root (U2R)

User-to-root attack is attempts by an unauthorized user to gain administrative privileges. The attacker starts outs with access to a normal user account on the system (perhaps gained by sniffing password, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

3.4. Remote-to-Local (R2L)

Remote-to-local attack is the kind of intrusion attack where the remote intruder consistently sends packets to a local machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

III. PROPOSED METHODOLOGY

The KDD Cup 99 Data set is defined data set and automatically used in the system by using some software. It has 41 features that are also pre-defined in the system. Intrusion detection system is basically works on this data set.

KDD data set has two sets: Training set and testing set. Firstly, the feature extraction is done with the help of correlation and feature selection is done with the help of KPCA. These are the methods of data mining. Then, the ANN is used for processing the features to find the accuracy, precision and recall. At last, the genetic algorithm is used as a classifier and performed the error detection for the neural networks.

The flow chart of the proposed system is in figure 1. The steps of the proposed system flow chart are as follows:

1. The KDD99 cup is huge database. So, it is not possible to load huge database. So, the KDD99 is dividing into two sets: trained set and tested set.

2. The training set is loaded and for that the features selection is done for various dimensions separately.

3. The correlation is used for feature selection and selected features id defined in the class {+1,-1}.

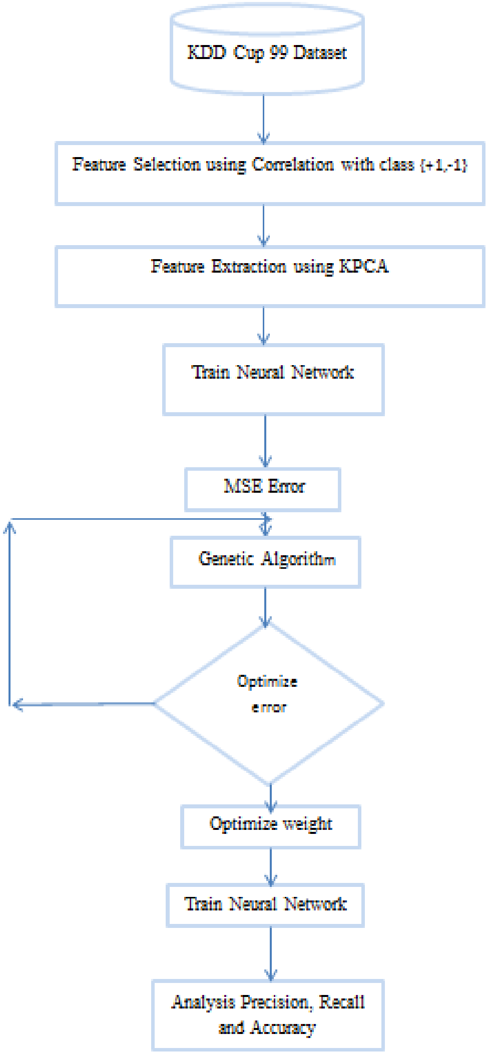
KDD Cup 99 Dataset
Feature Selection using Correlation with class {+1,-1}
Feature Extraction using KPCA
Train Neural Network
MSE Error
Genetic Algorithm
Optimize error
Optimize weight
Train Neural Network
Analysis Precision, Recall and Accuracy

Figure 1: System flow diagram

4. Then feature is extracted by KPCA. The feature extraction is done for the selected features. The feature extraction is done with the help of Eigen values and Eigen vectors.

5. Train neural network is used for input process for test set. The trained set generates input for the test set.

6. Neural network find some error for the input data and produce noise for test set.

7. The error is optimized by the genetic algorithm and if optimization is done properly(no error in the input data), it goes to optimize weight

8. All input data has some weight and that weight is optimized for error free data.

9.Again, the train neural is running and produces the test data set.

10. Test data is finally done with results and produce performance for the 10 fold values.

IV. RESULTS AND DISCUSSIONS

The performance of the proposed system is done with the help of Machine Learning - Confusion Matrix.
Confusion matrix is a 2×2 matrix, where the rows represent actual classes; while the columns have the values correspond to the predicted classes.

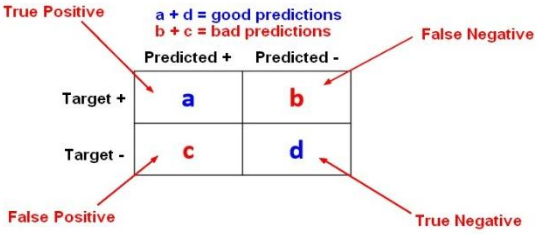

Figure 2: Confusion Matrix

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

*Precision = TP / (TP+FP)*100*

*Recall = TP / (TP+FN)*100*

*Acc. = TP+TN / (TP+TN+FP+FN)*100*

Experiment 1

Table 4.1: Features set for 15 features

Accuracy	Precision	Recall
100	100	100
83	50	83
100	100	100
83	50	83
92.5	50	92.5
59	50	59
100	100	100
92.5	50	92.5
83	50	83
92.5	50	92.5

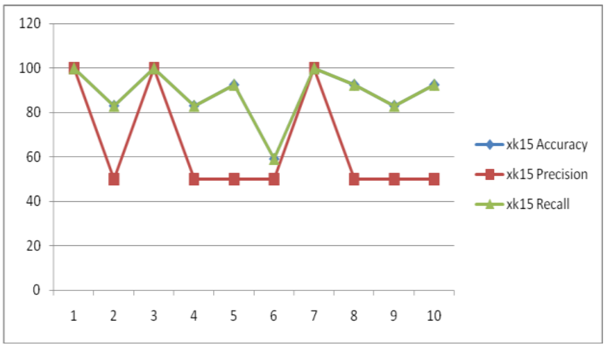

Figure 4.1: Graphical Representation of the performance of 15 features

- **Performance for 20 set of features:**
-

Table 4.2: Performance for 20 set of features

Accuracy	Precision	Recall
92.5	50	92.5
100	100	100
100	100	100
92.5	50	92.5
100	100	100
83	50	83
83	50	83
83	50	83
83	50	83
100	100	100

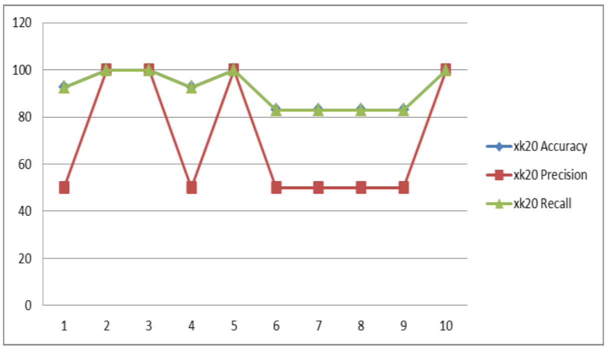

Figure 4.2: Graphical Representation of the performance of 20 features

- **Performance for 25 set of features:**

Table 4.3: Performance for 25 set of features

Accuracy	Precision	Recall
83	50	83
100	100	100
92.5	50	92.5
83	50	83
90.5	50	90.5
100	100	100
100	100	100
100	100	100
92.5	50	92.5
100	100	100

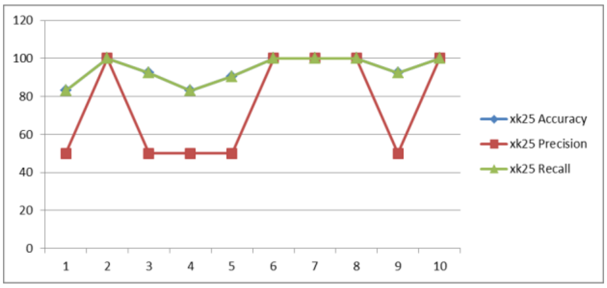

Figure 4.3: Graphical Representation of the performance of 25 features

- **Performance for 30 set of features:**

Table 4.4: Performance for 30 set of features

Accuracy	Precision	Recall
100	100	100
83	50	83
100	100	100
83	50	83
100	50	83
92.5	100	100
92.5	50	92.5
92.5	50	92.5
92.5	50	92.5
91.9	50	92.5

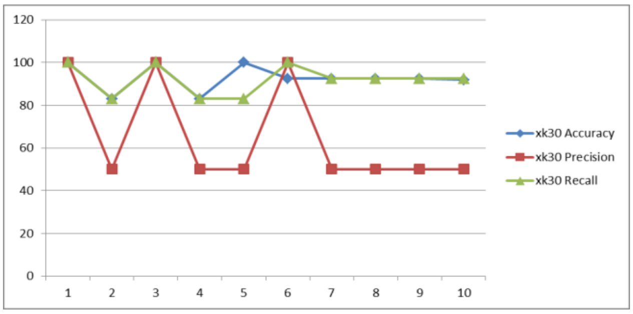

Figure 4.3: Graphical Representation of the performance of 30 features

These are the feature set values in tables and graphs as per the performance. Now, we combined the average value of the all the metrics of the data set and then perform the table and graph.

Table 4.5: Comparative Performance for different set of features

Set of features	Accuracy	Precision	Recall
‘15’	88.55	65	88.55
‘20’	91.7	70	91.7
‘25’	94.15	75	94.15
	92.79	65	91.9

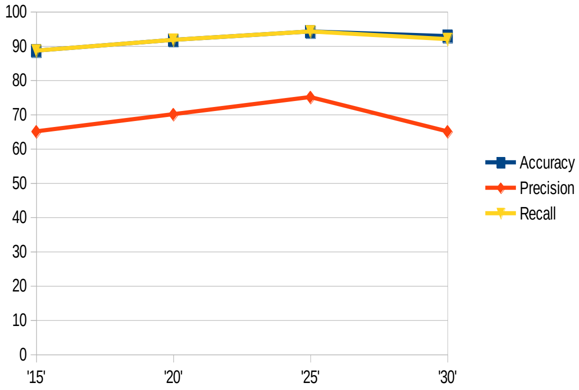

Figure 4.3: Graphical Representation of the performance of all set of features

The graph shows the accuracy, precision and recall for the 4 different set of features. And according to this graph the accuracy is almost good for all the features and if we are not using the different attacks of the KDD99 cup the accuracy is also good without that. And the accuracy and recall is almost same because the recall is almost like the accuracy because it is true values of the functions.

5. CONCLUSIONS AND FUTURE SCOPE

CONCLUSIONS: This research study the performance measures of IDS is important for the security purposes. KDD 99 has 41 features. The IDS approached is used with the help of neural network technique of data mining and the Genetic algorithm is used as a classifier. The all features of KDD99 are used in this study. In this research, the feature is selected and extracted instead of using all features. The feature selection is done with the help of Correlation and feature extraction is done with the help of KPCA. The selection of features is according to the Eigen values of the features. The neural network is used for the change the weightage of the error. The neural network is basically runs many times and change weightage according to this. Then it creates the confusion matrix for calculation the value of precision and recall. The genetic algorithm is used as a classifier for error detection. The test results proves the feature selection of IDS get improved the selected features are used alone instead of the all features. All features are produces the redundancy, complexity in the system and decreases the accuracy. But a selected feature increases the accuracy, precision and recall for all the features set.

FUTURE SCOPE: The proposed models can be checked on different Intrusion detection system, to establish the new benchmarks on Intrusion detection.In future work, new hybrid model for intrusion detection can be built by optimizing the different machine learning algorithms.

More parameters can be set for network features to improve the rate of intrusion detection, by further applying more techniques on proposed model this model can be laid as basic foundation for real life intrusion detection in future.

REFRERENCES

[1] Anup K. Ghosh, Schwartzbard A, Schatz M, "*Workshop on Intrusion Detection and Network Monitoring*," Santa Clara, California, USA.,1999.

[2] Kumar J,D, "*Attack Development for Intrusion Detection Evaluation Attack Development for Intrusion Detection Evaluation*", Massachusetts Institute of Technology,2000.

[3]Sabhnani, M, Serpen,G, "*Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context*," In Proc. of the International Conference on Machine Learning Models Technologic and Applications, pp 209-215,June 2003. \

[4]Sung A. and Mukkamala S., "*Identifying Important Features for Intrusion Detection using Support Vector Machines and Neural Networks*," in Proceedings of Inter-National Symposium on Applications and the Internet, pp. 209-217, 2003.

[5] Chebrolu S., Abraham A., and Thomas P., "*Feature Deduction and Ensemble Design of Intrusion Detection Systems*," Computers and Security, vol. 24, no. 4, pp. 295-307, 2005.

[6] Kayacik G., Zincir-Heywood N., and Heywood I., "*Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets*," in Proceedings of the 3rd Annual Conference on Privacy, Security and Trust, Andrews, Canada, 2005.

[7]Ren Hui Gong, Mohammad Zulkernine, Purang Abolmaesumi (2005) "*A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection*" in Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05).

[8] Lisong Pei, Schütte, J, "*Intrusion detection system*," July 2007, Carlos Simon.

[9] Jimmy Shun and Heidar A. Malki (2008) "*Network Intrusion Detection System Using Neural Networks," ICNC*, 2008, 2013 International Conference on Computing, Networking and Communications (ICNC), 2013 International Conference on Computing, Networking and Communications (ICNC) 2008, pp. 242-246.

[10] Mahbod Tavallaee,M, Bagheri, E, W, Lu, and Ali A,G(2009), "*A Detailed Analysis of the KDD CUP 99 Data Set*" IEEE Symposium on computational intelligence in security and defense application, pp. 1-6, 8-10 July 2009.

[11] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey(2010)*"Intrusion detection using data mining technique"* International conference on information retrieval and knowledge management; 2010. p. 200–4.

[12] Ahmed Youssef and Ahmed Emam (2011) "*Network Intrusion Detection Using Data Mining And Network Behaviour Analysis*" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[13] Virendra Barot and Durga Toshniwal (2012) "*A new data mining based hybrid network Intrusion Detection model*" IEEE 2012.

[14] Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas (2012) "*An Implementation of Intrusion Detection System using Genetic Algorithm*" International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

[15]G.V. Nadiammai and M. Hemalatha (2013) "*Effective approach toward Intrusion Detection System using data mining techniques*" Egyptian Informatics Journal (2014) 15, 37–50.

[16] Karan Bajaj and Amit Arora (2013) " *Improving the Intrusion Detection using Discriminative Machine Learning Approach and Improve the Time Complexity by Data Mining Feature Selection Methods*" International Journal of Computer Applications (0975 – 8887) Volume 76– No.1.

[17] Sharmila Kishor Wagh, Vinod K. Pachghare, Satish R. Kolhe(2013) "*Survey on Intrusion Detection System using Machine Learning Techniques*"International Journal of Computer Applications (0975 – 8887) Volume 78 – No.16.

[18] Jayveer Singh and Manisha J. Nene (2013) *"A Survey on Machine Learning Techniques for Intrusion Detection Systems"* International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11.

[19] E. Kesavulu Reddy (2013) "*Neural Networks for Intrusion Detection and Its Applications*" Proceedings of the World Congress on Engineering 2013 Vol II, WCE 2013, July 3 - 5, 2013, London, U.K.

[20] Mradul Dhakar and Akhilesh Tiwari (2013) "*A Novel Data Mining based Hybrid Intrusion Detection Framework*" Journal of Information and Computing Science Vol. 9, No. 1, 2014, pp. 037-048.

[21] Vishnu Prasad Goranthala, Naresh Goke, Ravi Kumar (2013) "*Analysis of Intrusion & Alert over the Computer Network Data Packet*" International Journal of Engineering and Innovative Technology Vol.2, No. 8, pp. 254- 257.

[22] Megha Aggarwal and Amrita (2013) "*Performance Analysis Of Different Feature Selection Methods In Intrusion Detection*" International Journal Of Scientific & Technology Research Volume 2, Issue 6.

[23] Mostaque Md. Morshedur Hassan (2013) "*Network Intrusion Detection System Using Genetic Algorithm and Fuzzy Logic*" International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 7.

[24] Salah Eddine Benaicha et.al (2014) "*Intrusion detection system using genetic algorithm*" Science and Information Conference (SAI), 2014. IEEE, 2014.

[25] Lucky Sharma (2014) "*New Hybrid Intrusion Detection System Based On Data Mining Technique to Enhance Performance*" International Journal of Computational Vol. 04, Issue 12.

[26] Vasim Iqbal Memon and Gajendra Singh Chandel (2014) "*A Design and Implementation of New Hybrid System for Anomaly Intrusion Detection System to Improve Efficiency*" Vasim Iqbal Memon et al Int. Journal of Engineering Research and Applications Vol. 4, Issue 5(Version 1), pp.01-07.

[27] L.Dhanabal and Dr. S.P. Shantharajah (2015) "*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.

[28] Dr. S.Vijayarani1 and Ms. Maria Sylviaa.S (2015) "*Intrusion Detection System – A Study*" International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1.

[29] Kaliappan Jayakumar et.al (2015) "*Intrusion Detection using Artificial Neural Networks with Best Set of Features*" The International Arab Journal of Information Technology, Vol. 12, No. 6A, 2015.

[30] Vinod Rampure and Akhilesh Tiwari (2015) "*A Rough Set Based Feature Selection on KDD CUP 99 Data Set*" International Journal of Database Theory and Application Vol.8, No.1 , pp.149-156.

[31] KDDCup99 Dataset., available at:http://kdd.ics.uci.edu//databases/kddcup99/kddcup99.html, last visited 2013.

[32]Kayacik G., Zincir-Heywood N., and Heywood I., "*Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets*," in Proceedings of the 3rd Annual Conference on Privacy, Security and Trust, Andrews, Canada, 2005.

[33] Chena Y., Abrahama A., and Yanga B., "Feature Selection and Classification using Flexible Neural Tree," *Journal of Neuro Computing*, vol. 70, no. 1, pp. 305-313, 2006.