

# An Efficient Neural Network Technique for Misuse Detection in IDS

Nishu Rooprai<sup>1</sup> Rupika Rana<sup>2</sup>

<sup>1</sup>M. Tech. Student <sup>2</sup>Assistant Professor

<sup>1,2</sup>Department of Computer Science & Engineering

<sup>1,2</sup>PTU Regional Centre, SSIET, Dera Bassi, India

**Abstract**— Machine learning methods for Intrusion detection system has been giving elevating accuracy and quality detection potential on novel strikes. At present the two primary techniques of recognition are signature-based and anomaly-based on their detection method. IDS occur in a collection of "spices" and approach the target of recognizing suspicious traffic in alternate methods. There are network based (NIDS) and host based (HIDS) intrusion detection systems. Artificial neural networks (ANN) have become very important and profitable in domains as pattern classification and regression, because using rule-based programming does not offer the right solution or any solution at all. The neural network contains both the supervised and as well as unsupervised learning. In this research work, a neural network based technique is proposed to prevent misuse detection in intrusion detection system. The technique will be implemented on labeled and non-labeled data and combination of both. The technique is implemented for accuracy of intrusion detection in each iterations of detection.

**Key words:** Neural Network, Intrusion Detection, Attacks, Misuse Detection, R Language

## I. INTRODUCTION

Intrusion detection is the approach of observing the events happening in a computer system or network and inspecting them for indications of feasible incidents, which are infringements or menacing ultimatums of violation of computer security strategies, acceptable use policies, or standard security practices. Intrusion prevention is the method of conducting intrusion detection and striving to stop detected possible incidents. Intrusion detection and prevention systems (IDPS) are especially concentrated on discovering possible incidents, logging information about them, seeking to cease them, and outlining them to security administrators. Furthermore, organizations operate IDPSs for alternate initiatives, such as identifying issues with security policies, documenting existing threats, and deterring individuals from violating security policies. IDPSs have become a mandatory addition to the security infrastructure of nearly every organization.

There are four types of IDPS technologies:

- Network-Based, which monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity
- Wireless, which monitors wireless network traffic and analyzes it to identify suspicious activity involving the wireless networking protocols themselves
- Network Behavior Analysis (NBA), which examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware, and policy

violations (e.g., a client system providing network services to other systems) [18]

- Host-Based, which monitors the characteristics of a single host and the events occurring within that host for suspicious activity.

## II. LITERATURE SURVEY

S. K. Wagh et al. [1] have proposed an algorithm for semi-supervised learning for intrusion detection. The suggested algorithm can significantly enhance the tuning of any base classifier in the absence of more labeled data. An experimental framework was illustrated in the paper in which various supervised and semi-supervised algorithms/ learning methods are evaluated in an intrusion detection system with same, small portion of labelled data. Through the experiments it was shown that the proposed SSL algorithm can perform better than the other traditional supervised algorithms and semi-supervised methods. The results confirmed that using accuracy on the original labelled data to further decide whether to accept the new unlabeled data (confidence data) into the next iterations or not is an effective way to improve the performance in semi-supervised learning. The observations suggested that semi-supervised learning was used to solve the problem of availability of the large amount labelled data.

S. K. Wagh et al. [2] proposed an algorithm for semi-supervised learning for intrusion detection using a boosting framework is proposed. The strength of the proposed algorithm lies in its ability to improve the performance of any given base classifier in the presence of unlabeled data. An experimental framework was proposed in which many supervised and semi-supervised learning methods can be evaluated in an intrusion detection system. The experiments demonstrated that the proposed algorithm had an excellent accuracy.

Initially, only 10,000 labeled data and 1, 00,000 unlabeled data were present. In the first iteration, as the training sets contain only labeled data, so proposed semi supervised algorithm worked like as supervised learning. In second iteration proposed algorithm worked as semi-supervised learning because after end of the first iteration unlabeled data (most confident data) has been added to the training file so it is a combination of labeled and unlabeled data. After third iteration training set had been risen up to 60000 data. Proposed semi supervised algorithm gave 99.516 % accuracy in the last iteration. Proposed algorithm worked better for Dos attack as compared to other attacks, as it was false positive rate is 0.102%.

G. V. Nadiammai et al. [3] integrated the data mining concept with an ID to identify the relevant, hidden data of interest for the user effectively and with less execution time. Four issues such as Classification of Data, High Level of Human Interaction, Lack of Labeled Data, and Effectiveness of Distributed Denial of Service Attack

were being solved using the proposed algorithms like EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPERAA Algorithm respectively. Proposed algorithm had been tested using KDD Cup dataset. All proposed algorithm had shown better accuracy and reduced false alarm rate when compared with existing algorithms.

It had detected 149 attacks out of 180 (83%) attacks after training in one week attack free traffic data. This approach helped to overcome the human interaction toward preprocessing. Regarding third issue, the proposed Semi-Supervised approach was 18.1% better than RSVM, 18.9% better than PCKCM, 19.9% better than the FCC. To solve the overwhelming problem of supervised and unsupervised methods, the semi supervised approach had been carried out. Finally, based on the mitigation of DDoS attack scenario, the port hopping concept was used depending upon the message length. Hence the message loss was greatly reduced and it did not create severe damage if happens. Both the security and performance measures with a variable clock drift mechanism had been evaluated. Experimental results proved that the proposed algorithm solved many defects and issues.

V. Mahajan et al. [4] worked on Network Traffic Classification which was an important process in various network management activities like network planning, designing, workload characterization etc. Network traffic classification using traditional techniques such as well-known port number based and payload analysis based techniques were no more effective because various applications used port hopping and encryption technique to avoid detection. Recently machine learning techniques such as supervised, unsupervised and semi supervised techniques were used to overcome the problems of traditional techniques. In this work authors used semi supervised machine learning approach and proposed distance based semi supervised clustering and probabilistic assignment technique for network traffic classification. This technique used only flow statistics to classify network traffic. It permitted to build the classifier using both labeled and unlabeled instances in training dataset.

The results showed that more than 92% precision achieved for all classes. Second, probe class achieved 100% precisions i.e. the instances belong to other class were not classified as belongs to probe class. Third, the normal class achieved lowest precision indicates that the other instances were misclassified as it was belonged to this class as compared to others. Forth, more than 95% recall achieved for all classes. Fifth, DoS class had achieved 97.5% recall i.e. the instances belong to this class were more correctly classified. Sixth, U2R class achieved lowest recall values i.e. the large number of instances belongs to this class are misclassified as compared to the other classes.

Chien-Yi Chiu et al. [5] studied that Intrusion Detection Systems (IDSs) had been deployed in computer networks to detect a wide variety of attacks were suffering how to manage of a large number of triggered alerts. Thus, reducing false alarms efficiently had become the most important issue in IDS. In this paper, authors introduced the semi-supervised learning mechanism to build an alert filter, which had reduced false alarms up to 85% and still had kept a high detection rate. In our semi-supervised learning

approach, authors only needed a very small amount of label information. This would save a huge security officer's effort and make the alert filter be more practical for the real systems. Numerical comparison with conventional supervised learning approach with the same small portion labeled data, this method had significantly superior detection rate as well as in the false alarm reduction rate.

In this paper, authors used a semi-supervised algorithm, Two-Teachers-One-Student (2T1S), as the learner of our machine learning based analysis engine. 2T1S was a multi-view algorithm. Different from regular multi-view methods, 2T1S selects different viewed in the feature space rather than in the input space. 2T1S elegantly blended the concepts of co-training and consensus training. Through co-training, the classifier generated by one view could "teach" other classifiers constructed from other views to learn, and vice versa; and by consensus training, predictions from more than one view could give us higher confidence for labeling unlabeled data. In practice, given three different views, 2T1S selected two views as teachers for consensus training and the remaining view as the co-training partner. The classification had answered from two classifiers (two teachers) represent the consensus result, which was used to teach the third view (the student) to learn the labels for unlabeled data. This process was performed for each choice of teachers-student combination. After the student had learned the data, the newly learned labeled data was added to the student's original labeled data set, as the set of guessed labeled data can be included for training in the next step if it was part of the teachers' sets in the next step. The whole process was run iteratively and alternately until some stopping criteria were satisfied.

C. Chen et al. [6] had worked on research interests of applying or developing specialized machine learning techniques that had attracted many researchers in the intrusion detection community. Existing research work showed that the supervised algorithms deteriorated significantly if unknown attacks were present in the test data; the unsupervised algorithms exhibit no significant difference in performance between known and unknown attacks but their performances were not that satisfying. In this paper, authors investigated the capabilities of semisupervised learning methods, both semi-supervised classification methods and semi-supervised clustering methods, for intrusion detection. Authors main contributions were: Firstly, two semi-supervised classification methods, Spectral Graph Transducer and Gaussian Fields Approach, were proposed to take the task of detecting unknown attacks. A special data set was designed to test the capabilities of all classification methods (both supervised and semi-supervised versions) for detecting unknown attacks. Experiments of comparing them with other seven traditional supervised learning methods (eleven versions) were presented. Results showed that their performances were much better than those of the other seven traditional supervised learning methods on detecting unknown attacks. Secondly, one semi-supervised clustering method—MPCK-means was introduced to improve purely unsupervised clustering methods on intrusion detection. Experiment was presented the comparison between two versions of MPCK-means and unsupervised learning method—K-means. The result showed that performance of MPCK-means (both two

versions) was much better than that of K-means. Finally, after analyzing the results of experiments carefully and deeply, proposed a potential learning method—transfer learning, which might be more adept in detecting unknown attacks with the aid of accumulated records (training examples) of known attacks.

### III. METHODOLOGIES

#### A. K2 Learning

The K2 algorithm (Cooper and Herskovits 1992) uses a greedy search and may impose no restriction on the number of parents a node has. The K2 search begins by assuming that a node (representing a discrete variable) has no parents and then adds incrementally that parent from a given ordering whose addition increases the score of the resulting structure the most. We stop adding parents to the node when the score stops to increase.

Inference in BNs is the task of calculating the conditional probability distribution of a subset of the nodes in the graph (the “hidden” nodes) given another subset of the nodes (the “observed” nodes). In a classification problem a hidden node represents the class variable, the observed nodes represent the features, and inference is conducted. (2). Use the junction tree algorithm for inference (Huang and Darwiche 1994; Lauritzen and Spiegelhalter 1988).

#### B. Bayesian

A Bayesian Network is a graph in which

- 1) A set of random variables makes up the nodes in the network.
- 2) A set of directed links or arrows connects pairs of nodes.
- 3) Each node has a conditional probability table that quantifies the effects the parents have on the node.
- 4) Directed acyclic graph (DAG), i.e. no directed cycles

#### C. KDD Cup Dataset

All experiments are carried out with KDD CUP 1999 data set. In spite of many drawbacks mentioned, it is used as reliable benchmark data set for many researches on network based intrusion detection algorithms. KDD Data set contained labeled TCP/IP packets. Each packet is having 41 features, some of which are nominal attributes and others are numerical attributes. So it avoids the task of “Feature extraction” and “Data labeling”. Hence attention can be concentrated on the efficiency and preciseness of the algorithm.

#### D. R Tool

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

R and its libraries implement a wide variety of statistical and graphical techniques which includes linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R

itself, which makes it easy for users to follow the algorithmic choices made.

#### E. Proposed Methodology

The following strategy will be followed to get the desired results.

- Step 1: Implement all three existing techniques using K2, Bayesian and Datasets.
- Step 2: Select the database and perform learning as well as testing in R tool.

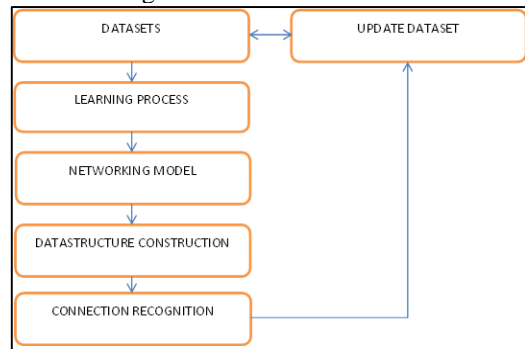


Fig. 1: Framework for adaptive intrusion detection

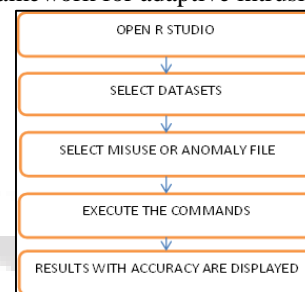


Fig. 2: Framework for ID with RStudio

### IV. RESULTS

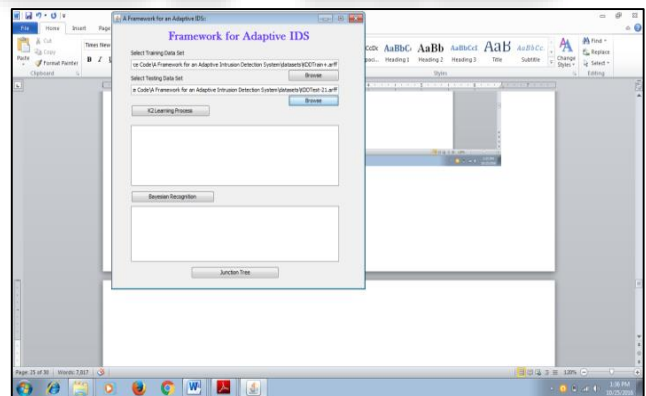


Fig. 3: GUI of loading of training and testing dataset

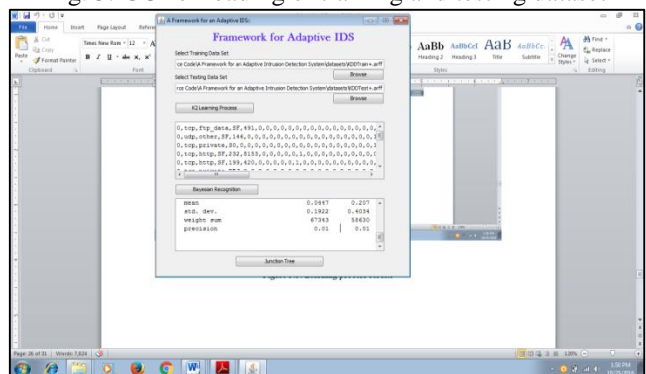


Fig. 4: Networking through Bayesian for database 1

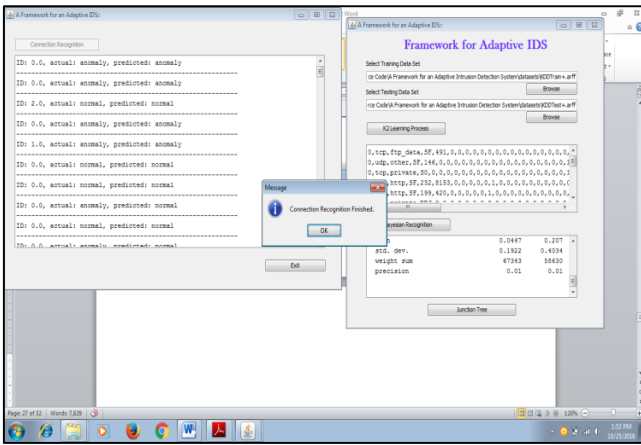


Fig. 5: Data structure creation for database 1

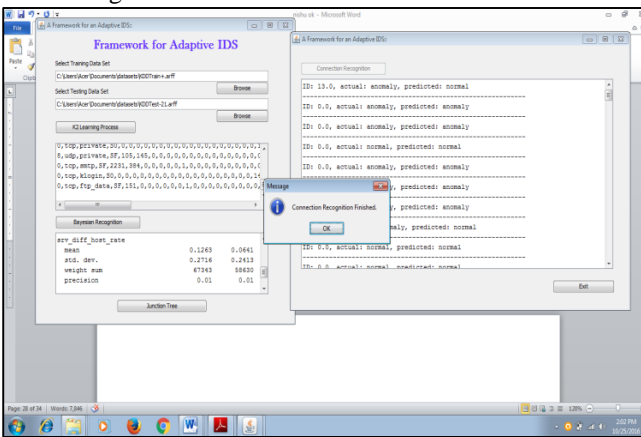


Fig. 6: Data structure creation for database 2

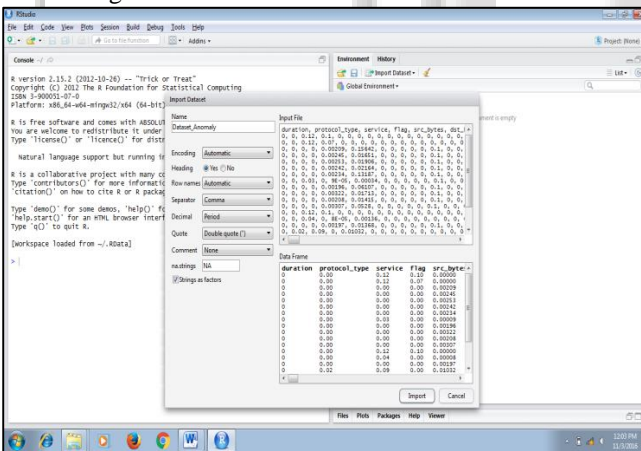


Fig. 7: Importing anomaly dataset in RStudio explorer

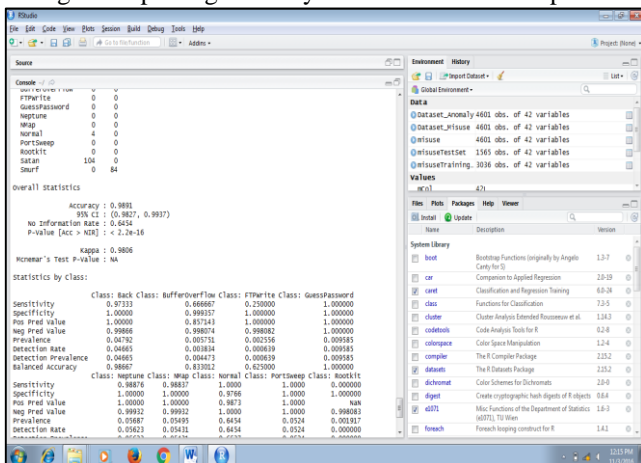


Fig. 8: Displaying results of anomaly dataset

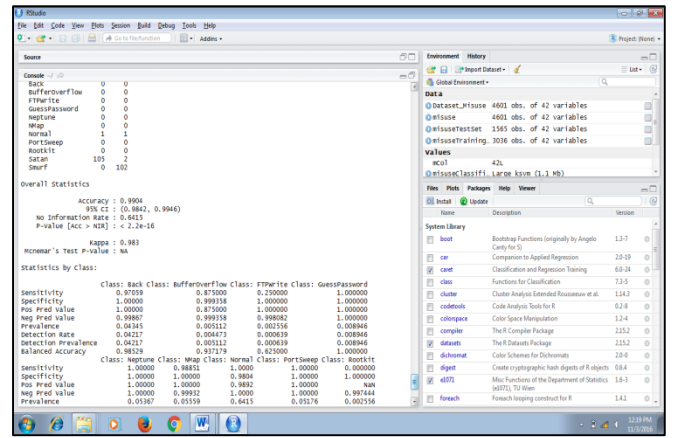


Fig. 9: Displaying result of misuse dataset

Classification Approach	Accuracy for Anomaly based detection	Accuracy for Misuse based detection
Using nueralnet package in R	99%	98%

Table 1: Accuracy Summary

## V. CONCLUSION

The results drawn are obtained from running tests for misuse and anomaly approaches. The original dataset had with 41 attributes. Total 10 types of attack are used and the output of each is in the form of a confusion matrix. The accuracy of the overall system is 99 and 98 percent for anomaly and misuse based signature detection.

In future more number of attacks can be taken. Also number of attributes can be increased for both anomaly and misuse based detection. The work can also be done on the server so that maximum detection of intrusions is possible.

## REFERENCES

- [1] S. K. Wagh and S. R. Kolhe, "Effective semi-supervised approach towards intrusion detection system using machine learning techniques", Int. J. Electronic Security and Digital Forensics, Vol. 7, No. 3, pp. 290-304, 2015.
- [2] S. K. Wagh and S. R. Kolhe, "Effective Intrusion Detection System Using Semi-Supervised Learning", IEEE ICDIMC, pp. 1-5, 2014.
- [3] G. V. Nadiammam, M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal. Vol. 15, Issue 1, pp 37–50, March 2014.
- [4] V. Mahajan, B. Verma, "Implementation of Distance Based Semi Supervised Clustering and Probabilistic Assignment Technique for Network Traffic Classification", International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 2, pp.1249-1252, Mar-Apr 2012.
- [5] Chien-Yi Chiu, Yuh-Jye Lee, Chien-Chung Chang, Wen-Yang Luo, and Hsiu-Chuan Huang, "Semi-supervised Learning for False Alarm Reduction", P. Perner (Ed.): ICDM 2010, LNAI 6171, pp. 595–605, Springer-Verlag Berlin Heidelberg 2010.
- [6] Chuanliang Chen, Yunchao Gong, and Yingjie Tian, "Semi-Supervised Learning Methods for Network

- Intrusion Detection”, 1-4244-2384-2/08/\$20.00 c\_ 2008 IEEE.
- [7] P. Laskov, P. Dussel, C. Schafer, et al., “Learning Intrusion Detection: Supervised or Unsupervised,” In Proc. of Image Analysis and Processing - ICIAP 2005, 13th International Conference, pp. 50-57, 2005.
- [8] J. McHugh, A. Christie, J. Allen, “Defending yourself: The role of intrusion detection systems,” IEEE Software, pp. 42–51, Sept./Oct. 2000.
- [9] J. Aslam, S. Bratus, V. Pavlu, “Semi-supervised Data Organization for Interactive Anomaly Analysis,” In Proc. of the 5th International Conference on Machine Learning and Applications, pp. 55-62, 2006.
- [10] X. Zhu, “Semi-supervised learning literature survey,” Tech. Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.
- [11] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen and Carey William Son, “Semi-Supervised Network Traffic Classification”, in proc. of ACM SIGMETRICS’07, San Diego, California, USA, vol.35, June 2007, 369-370.
- [12] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen, and Carey Williamson, Offline/Realtime Traffic Classification Using Semi-Supervised Learning, Technical Report, Department of Computer Science, University of Calgary, February 2007. (Manuscript 22 pages)
- [13] Chuanliang Chen, Yunchao Gong and Yingjie Tian, “Semi-Supervised Learning Methods for Network Intrusion Detection”, in proc. of IEEE International Conference on Systems Man and Cybernetics (SMC), 2008, 2603-2608.
- [14] Levi Lelis and Jorg Sander, “Semi-Supervised DensityBased Clustering”, in proc. of Ninth IEEE International Conference on Data Mining , Miami, FL , Dec 2009, 842-847.
- [15] Liu Bin and Tu Hao, “An Application Traffic Classification Method Based on Semi-Supervised Clustering”, A 2nd International Symposium on Information Engineering and Electronics Commerce (IEEC), 2010, 1-4.
- [16] Amita Shrivastav and Aruna Tiwari, “Network Traffic Classification using Semi-Supervised Approach”, Second International Conference on Machine Learning and computing (ICMLC), 2010, 345-349.
- [17] KDD CUP 1999 dataset available at <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.
- [18] K. Scarfone and P. Mell, “Guide to Intrusion Detection and Prevention Systems,” NIST Special Publication 800-94, 2007.
- [19] A. Lazarevic, L. Ertöz, V. Kumar, et al, “A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection,” In Proc. of the 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1-3, 2003.
- [20] Successful Real-Time Security Monitoring, Riptech Inc. white paper, Sep. 2001.