# AN EFFICIENT NETWORK INTRUSION DETECTION BY ENSEMBLE LEARNING

**M.Deepa[1], Dr.P.Sumitra[2],**
[1]Ph.D Research Scholar,
Legithasai2010@gmail.com
[2]Professor,
sumitravaradharajan@gmail.com
PG and Research Department of Computer Science and Applications,
Vivekananda College of Arts and Sciences for
Women(Autonomous),Elayampalayam,Tiruchengode,Namakkal(DT)TamilNadu,India

## ABSTRACT

In the globe the corners of all communication trade are connected together by using advance network technology. At the same decade intruders are more effectively make attacks on the networks. Most of the intrusion detection system are developed by using single as well as hybrid algorithms but the key point is selecting the appropriate features on the dataset because the proper feature selection yields a high accuracy and reduce the false positive rate. In this paper an ensemble learning approach are introduced. The NSL-KDD dataset are habitually used in this field of intrusion detection system. The NSL-KDD dataset are preprocessed with attribute selection algorithms and the random forest algorithm by selecting the preferred features.

## KEYWORDS

IDS, Feature Selection, classification, WEKA, Machine Learning

## I.INTRODUCTION

In the digital era, all the information is transferred through network using newest technologies. In the meanwhile, the confidentiality of communication is very deprived because many vulnerable activities are increased. The existing security policy like firewall doesn't preventing such types of hacks because of software application contains hidden vulnerability. The software application called Intrusion Detection System (IDS) monitors all unauthorized activities on the network. The IDS comes in many 'flavors' but it aims is to detecting suspicious activities. In this paper we are boosting the data mining for better accuracy. The NSL-KDD dataset is taken for implantation in WEKA environment. In this paper the performance of various data mining techniques are compared based on different parameters like time required, size of the tree, accuracy, kapha statistics, false positive obtained by various algorithms.

## II DATASET AND PREPROCESSING

### a) DATASET Description

For analyzing the efficiency of the algorithms, we have chosen NSL-KDD dataset. It is the inherit version of KDD CUP99 dataset. It is the good dataset for network because it reduces the irrelevant information from KDDCUP99 dataset. The NSL-KDD dataset consists of 42 attributes and 24 different types of attacks and these attacks are grouped into 4 categories. They are DoS, Probe, U2R and R2L. The original NSL-KDD dataset divided into training set and testing set. The training dataset consist of 25193 instances along with 13449 instances are normal data and 11744 are attack. The testing dataset consist of 2152 normal instances and 9698 instances are attacks.

**Dos:** Denial of service attack. In this type the accessing service of legitimate user is denied. The intruder act as a legitimate user.

**Probe:** In this type, the programs automatically examine the open network and access the IP address.

**R2L:** In this type the remote user act as an local user for accessing the local network.

**U2R:** In this type the local user trying to access the privilege for server system.

**Table 1: Feature List in NSL-KDD dataset**

| **Duration** | **su_attempted** | **same_srv_rate** |
|---|---|---|
| protocol_type | num_root | diff_srv_rate |
| Service | num_file_creation | srv_diff_host_rate |
| Flag | num_shells | dst_host_coTunt |
| src_byte | num_access_file | dst_host_srv_count |
| dst_byte | num_outbound | cmds |
| dst_host_same_srv_rate | land | is_host_login |
| dst_host_diff_srv_rate | wrong_fragment | is_gust_login |
| dst_host_same_src_port_rate | urgent | count |
| dst_host_srv_diff_host_rate | hot | srv_count |
| dst_host_serror_rate | num_failed_login | serror_rate |
| dst_host_srv_serro_rate | logged in | srv_serror_rate |
| dst_host_rerror_rate | num_compromised | rerror_rate |
| dst_host_srv_rerror_rate | root_shell | srv_rerror_rate |
| class | | |

**Table2: Attacks with their relevant group**

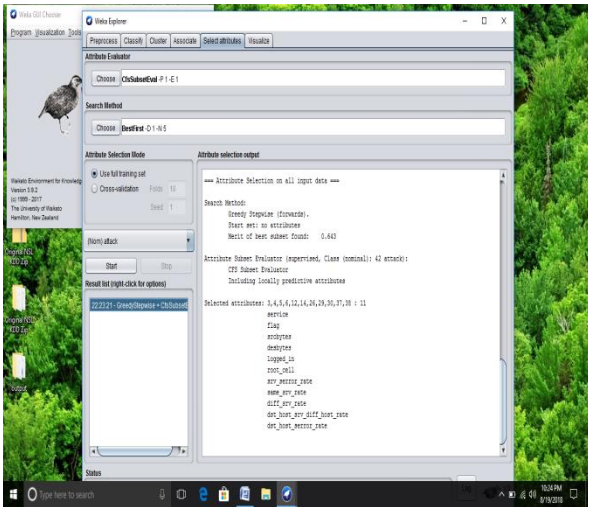
| **Attack group** | **Attacks** |
|---|---|
| DoS | Back, Neptune, Pod, Smurf, teardrop, Mailbomb, Processtable, Udpstorm, Apache2, Worm |
| Probe | Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint |
| R2L | Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Xlock, xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named |
| U2R | Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps |

### b) Preprocessing

Preprocessing plays a vital role in data mining process. Preprocessing is the first step for all data mining process. In preprocessing the redundant data, data inconsistency and noisy information are removed by applying the proper feature selection algorithm. The Real world data contains many redundant data and noisy data. The NSL-KDD data set is a real world data set. So before implementing data mining algorithms the data set are preprocessed.

**Feature selection**

Feature selection is critical to build a good intrusion detection system for several reasons. The feature selection not only improves the quality of the model but also makes the process of model in more efficient. Many algorithms do such types of work among these algorithms cfssubset evaluator are used for preprocessing. In this paper, the ability of every feature are evaluated individually along the degree of redundancy between these attributes. After that the subset are derived from original dataset based on the predictive ability of attributes. The resultant subsets are shown below. Only 11 attributes are selected among 42 attributes.



Attribute Evaluator: CfsSubsetEval -P 1 -E 1

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) attack

Result list (right-click for options): 22:23:21 - GreedyStepwise + CfsSubsetE

Attribute selection output

```
=== Attribute Selection on all input data ===

Search Method:
	Greedy Stepwise (forwards).
	Start set: no attributes
	Merit of best subset found:    0.643

Attribute Subset Evaluator (supervised, Class (nominal): 42 attack):
	CFS Subset Evaluator
	Including locally predictive attributes

Selected attributes: 3,4,5,6,12,14,26,29,30,37,38 : 11
                     service
                     flag
                     srcbytes
                     desbytes
                     logged_in
                     root_cell
                     srv_serror_rate
                     same_srv_rate
                     diff_srv_rate
                     dst_host_srv_diff_host_rate
                     dst_host_serror_rate
```

Waikato Environment for Knowledg
Version 3.9.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

**Fig 1: Feature selection(selected features)**

## III. METHODOLOGY

The proposed Intrusion detection system is designed by analyzing the machine learning algorithms and combines this prediction from the algorithm together to improve the stability and accuracy of algorithm than any individual model. In this paper an ensemble based techniques are used by heighten the result of the predictions. Bagging is an acronym for bootstrap aggregating. The algorithm bagging will take the training dataset as input and specify the number of iterations. Apply every algorithm and obtain the hypothesis (classifier). Analyze the vote for each algorithm and take the algorithm with high vote. In this paper the intrusion detection system are designed by using the ensemble classifier (HITNB). The good intrusion detection system have low false positive rate. The HoEffiding Induction tree algorithm produce better accuracy than the naïve bayes algorithm but the naïve bayes produce low false positives. After getting the result from every algorithm the results are analyzed and combine the results to boost up the performance of algorithms. Before implementing the algorithms the dataset is preprocessed. In preprocessing the unnecessary or irrelevant information are ignored.

### Naïve Bayes(NB)

Naïve bayes classifier is most broadly used for detecting attacks on the network. The NSL-KDD dataset have colossal quantity of data. The NB classifier provides the better accuracy for large data sets.

### Steps in NB

Step 1: Read the training dataset T
Step 2: calculate mean and standard deviation for each predictor.
Step 3: Repeat step 2 until conniving the probability of all prophet
Step 4: save the furthermost probability prophet

### Hoeffding tree

Hoeffding tree is a streaming decision tree based algorithm. It is an Induction Tree algorithm.

### Steps in Hoeffding

Step 1: Input the traning examples in (x,y) format.
Step 2: calculate gain value for each attribute.
Step 3: Build the tree using attributes with highest gain ratio.

### Proposed approach (HITNB)

Our proposed ensemble classifier hybridized the predictors from Hoeffding and naïve bayes classifier.

### Steps in HITNB

Step 1: Input the NSL-KDD data set.
Step 2: select the significant feature by preprocessing
Step 3: Build the HITNB classifier
Step 4: classify the network traffic as normal or attacks

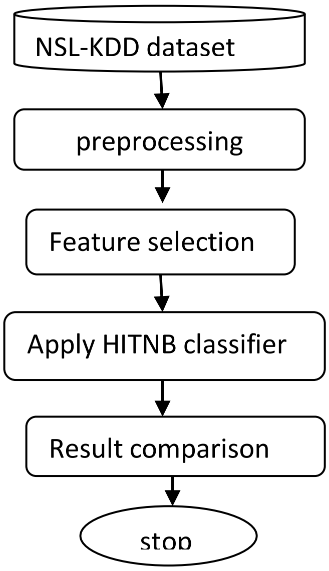


**Fig 2: Proposed IDS model**

## III. Experiment and Result analysis

In our experiment, the algorithms are implemented in weka 3.9.2 environment. Weka embraces several machine learning algorithms and visualization tools for data analysis and building efficient model.

| Parameter | Hoeffding | Naïve Bayes | HITNB |
|---|---|---|---|
| Correctly classified instances | 23006 | 20589 | 24865 |
| Incorrectly Classified Instances | 2187 | 4604 | 328 |
| Kappa statistic | 0.9371 | 0.8584 | 0.93561 |
| Mean absolute error | 0.0053 | 0.0132 | 0.0042 |
| Root mean squared error | 0.0457 | 0.0967 | 0.0387 |
| Relative absolute error | 2.31% | 5.75% | 2.28% |
| Time taken to build model | 1.08 seconds | 1.31seconds | 0.42 seconds |

There are several measurement metrics are available for measuring the attack detection. The most important metrics are

1. True positive(TP): corresponds to the attack is correctly identified as attack
2. True negative(TN):corresponds to the normal is correctly identified as normal
3. False Positive(FP): corresponds to the normal is incorrectly identified as attack
4. False Negative(FN): corresponds to the attack is incorrectly identified as normal.

| ALGORITHM | TP | TN | FP | FN | ACCURACY |
|---|---|---|---|---|---|
| Hoeffding | 10667 | 13047 | 29 | 100 | 88.447 |
| Naïve Bayes | 9667 | 12607 | 842 | 2069 | 94.1331 |
| HITNB | 11647 | 13420 | 14 | 97 | 97.48 |

Table 3 shows the accuracy rate of each algorithm. The accuracy of Hoeffding is 88.4%, the accuracy rate of Naïve bayes is 94% and the accuracy of HITNB is 97.48%. As a final point the proposed ensemble classifier HITNB took premier accuracy compared with the existing classifiers.

**Table 3:** Accuracy measurement parameters

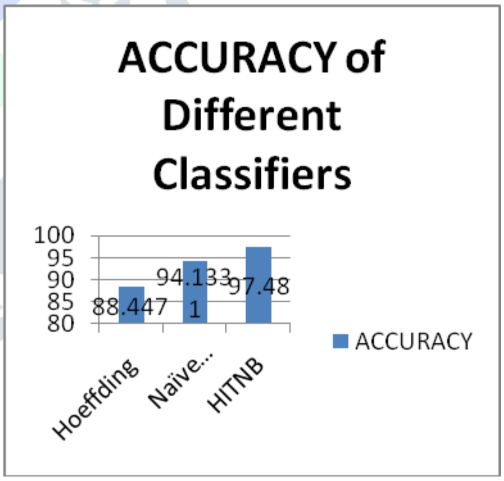


**Fig 3: Column chart depicts the existing and proposed algorithm**

**Table 4: Comparison of various metrics**

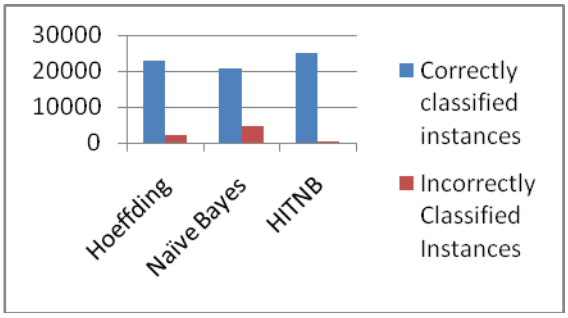


**Fig 4 : Comparison of correctly classified instances for various approaches**

## CONCLUSION

In this research work, we discuss an ensemble classifier (HITNB) for improving the performance of intrusion detection system. The proposed classifier professionally classifies the traffic on network as normal and attacks. Based on the result the proposed classifier is better than the existing individual classifiers. Each and every algorithm are implemented on NSL-KDD dataset in Weka. The result of these algorithms are compared with each other by using different parameters like True positive, correctly classified instances, kappa statistics, relative absolute error, relative mean square error and so on. From this analysis no individual algorithms produce better result so an ensemble classifier produces the better performance.

## REFERENCES

1. R. Mitchell and I.-R. Chen, “Behavior-rule based intrusion detection systems for safety critical smart grid applications,” *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1254–1263, Sep. 2013.
2. J. Valenzuela, J. Wang, and N. Bissinger, “Real-time intrusion detection in power system operations,” *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1052–1062, May 2013.
3. Ketan Sanjay Desale, Chandrakant Namdev Kumathekar and Arjun Pramod Chavan, “Efficient Intrusion Detection System using Stream Data Mining Classification Technique”, IEEE International Conference on Computing Communication Control and Automation ,2015.
4. Manish Kumar, Dr. M. Hanumanthapaa, “Intrusion Detection System using Stream Data Mining and Drift Detection Method”,4th lCCCNT - 2013 July 4-6,2013, Tiruchengode,India.
5. Bifet, Albert. "Mining Big Data in Real Time", Informatica37, pp:15-20, 2013.
6. Amreen Sultana, and M.A.Jabbar,” Intelligent Network Intrusion Detection System using Data Mining Techniques” , IEEE 2016.
7. G V Nadiammai, "Effective approach towards intrusion detection sytem using data mining techniques", Egyptian Informatics Journal, 15, pp 37- 50(2014).
8. Preeti Aggarwal, Sudhir Kumar Sharma, "Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection". 3rd International Conference on Recent Trends in Computing 2015 .
9. Mrutyunjaya panda et.al, "A hybrid intelligent approach for network intrusion detection", Procedia Engineering,45,pp 1-9(2012).
10. DikshantGupta, SuhaniSinghal, Shamita Malik and Archana Singh,” Network Intrusion Detection System Using various data mining techniques “,International Conference on Research Advances in Integrated Navigation Systems (RAINS - 2016), April 06-07, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India.
11. Revathi, S., and A. Malathi."A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection."International Journal of Engineering Research and Technology.Vol. 2.No. 12 (December- 2013).ESRSA Publications, 2013.
12. Salah Eddine Benaicha, Lalia Saoudi, Salah Eddine Bouhouita Guermeche, and Ouarda Lounis, “ Intrusion Detection System Using Genetic Algorithm “,Science and Information Conference 2014 August 27-29, 2014 | London, UK.