Enhancing the Performance of Network Intrusion Detection System by Combining Naïve Bayes, Decision Tree and K-Nearest Neighbors Algorithms

Abeselom Befekadu
Lecturer, Assosa University
Assosa, Ethiopia

ABSTRACT
Protecting the hostile network environment is a very difficult task. Although, there is no way to protect the network for hundred percent accuracy, so many researches tried to achieve the best security mechanisms for long time. Among the security mechanisms, network intrusion detection system is one of the well-known. The performances of the network intrusion detection systems that are developed have produce so many false alarm. To improve this false alarm rate this research combines three algorisms which are Naïve Bayes, Decision Tree and k-NN. The results found from the experiment showed that the combined algorithm improve the accuracy of the network intrusion detection system by up to 5%.

General Terms
Security, Algorithm

Keywords
Network Intrusion, Network Security, Intrusion Detection System

1. INTRODUCTION
Because of the increasing number of attack on computer network infrastructure; the importance of network security is not disputable. To secure the network infrastructure, intrusion detection systems have been the best option on the market. However, because of different issues surrounding the current intrusion detection system they lack the capacity to deter threats which comes from the network.

Different scholars tries to achieve the best performing intrusion detection system by using different kind of algorithms. But, the accuracy of the system still needs improvement to catch up with the current state of security.

To alleviate the problems surrounding the current intrusion detection system, this research proposes the combined algorithm which are Naïve Bayes, Decision Tree and K Nearest Neighbor to develop the network intrusion detection system.

1.1.Statement of the Problem
Network intrusion detection system is well known for its effectiveness on securing the network environment. One of the main challenge for intrusion detection system is the ability to detect intrusion during the time in which attackers try to manipulate the network for the purpose of stilling confidential data. But, the current intrusion detection system lacks the accuracy of detecting attacks.

To overcome this problem this research combines three algorithms. More specifically the following questions has been addressed in this research.

1. How to detect network based intrusions?
2. How to combine the three algorithms which are Naïve Bayes, Decision Tree and K Nearest Neighbor?
3. How to alert the network administrator at the time of attack.

1.2.Objective
1.2.1. General Objective
The main objective of this research is to develop network based intrusion detection system by combining three algorithms.

1.2.2. Specific Objectives
The specific objectives of this research are:

- To conduct a detailed literature review in order to understand the application domain for machine learning in network based intrusion detection system.
- To select the best detection algorithms and combine them to examine unknown attacks.
- To develop network based intrusion detection system by using the selected algorithms.
- To test and evaluate the combined intrusion detection system.

1.3.Scope and Limitation of the Study
The proposed integrated system has the following aspects:

- Identification of possible intrusions.
- Report generation about the intrusions.

The proposed system have the following limitations:

- The research only focuses on the intrusions that are caused by network based vulnerabilities.
- The designed solution is not 100% accurate. There is false positive and false negative in the result.
- The research only considered attacks which is found on NSL-KDD datasets.

2. RELATED WORKS
Sumaiya Thaseen, Ch. Aswani Kumar et al, [1] proposed dissimilar tree based categorization algorithms that arrange network events in intrusion detection systems and the examination is performed utilizing NSL-KDD 99 dataset. Dimensionality of the component of the dataset is lessened. The results demonstrate that Random Tree show grasps the most noteworthy level of rightness and decreased false alarm rate. This model is computed with other leading intrusion detection models to close its better predictive accuracy.

Nidhi Srivastav, Rama Krishna Challa et al, [2] show the layered structure incorporated with neural network to build a powerful and productive intrusion detection system. Such framework has tried different things with Knowledge Discovery and Data Mining (KDD) 1999 dataset. This framework was contrasted and displayed methodologies of intrusion detection which either employments neural network or in view of layered structure. The outcomes outline that the proposed strategy has high detection identification accuracy and less false alarm rate.

Mrutyunjaya Panda et.al, [3] proposed hybrid smart decision advancements utilizing data filtering by including guided learning techniques alongside a classifier to settle on more classified decisions together to identify network attacks. It is seen from the outcomes gotten that the Naive Bayes display is very engaging as a result of its uprightness, elegance, strength and viability. Then again, decision trees have demonstrated their effectiveness in both generalization and detection of new attacks. The outcomes demonstrate that there is no single best algorithm to beat others in all circumstances. In specific cases there may be reliance on the attributes of the data. To pick an appropriate algorithm, a domain expert or expert system may utilize the aftereffects of the classification with a specific end goal to settle on better decisions.

Juan Wang et.al, [4] introduced an intrusion detection system based on decision tree technology. In the procedure of building intrusion rules, data gain ratio is utilized as a part of information gain. The analysis outcome demonstrate that the C4.5 decision tree is possible and compelling, and has a high accuracy rate. His exploratory examination demonstrates that the C4.5 choice tree is a compelling method for the usage of decision tree and it gives right around 90% of classifier accuracy. Be that as it may, in this approach the error rate remains the same.

Yonav Freund et.al, [5] proposes a substituting decision tree with boosting. The new learning algorithm joins boosting and decision trees. In their paper they looked at the alternating decision tree with the C5.0 algorithm. On smaller datasets ADtree rapidly fits the information and ADtree comes to a small error after 50 iterations while the error of the stump support stays huge even after 200 cycles. This is a case in which huge limit of ADtree gives it leverage. Contrasting with the size of classifiers in everything except three cases the classifiers produced by the ADtree are considerably smaller than those produced by C5.0 by boosting. The error performance of this algorithm is near that of C5.0 with boosting.

3. DESIGN OF THE PROPOSED SYSTEM

The existing network intrusion detection systems are vulnerable to cyber-attacks. To alleviate this problem there is a need for effective algorithms which can identify intruders. This research proposes a more accurate intrusion detection system.

3.1.Architecture of the Proposed Intrusion Detection System

The network based intrusion detection system have several components and functions. The architecture of the proposed system is depicted in the following figure.

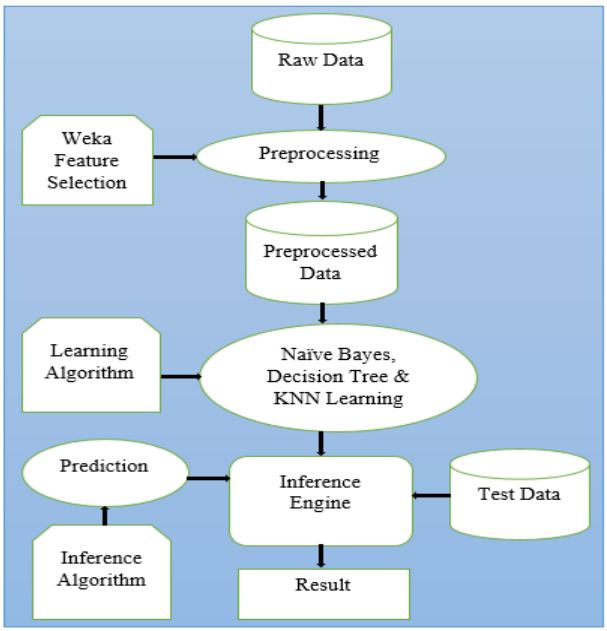

Figure 1: Architecture of the proposed integrated system

3.2.Network Intrusion Detection System Components

1. Raw Data

The input data for training phase is, the offline dataset which is found on the web for educational research purpose. It is labeled dataset that can be easily learnt by the system. For this case, the research has used simulated dataset called NSL-KDD for this phase. This dataset is selected because it is the latest version of all simulated dataset in the area of network security; redundant records are eliminated from training set and it is affordable to use for experiment purpose as it consists of reasonable number of instances both in the training and testing set [6].

2. Preprocessing

Data preprocessing is required to remove unwanted attributes from the dataset and build a dataset for Naïve Bayes and Decision Tree algorithms. Dataset feature extraction will be analyzed based on the attacks nature and extra domain information. The preprocessing phase is responsible for preparing the NSL-KDD dataset for the next phase which is Naïve Bayes and Decision Tree learning process. In this research WEKA attribute filtering has been used with other pre-processing techniques. The operations such as attribute selection, attribute filtering and instances filtering is applied in this phase. Those techniques improve the efficiency of the algorithm to classify the data correctly.

3. Naïve Bayes, Decision Tree And K-Nearest Neighbors Learning

Today's network environment is very crowded and uncertain. Detecting intrusions in this uncertain network environment is very difficult. Finding the best algorithm is a very challenging task in developing network based intrusion detection system. This research compare and contrasts every algorithms that are used to develop intrusion detection system and finds the combination of Naïve Bayes, Decision Tree and K- Nearest Neighbor algorithms are the best way to develop the system. Naïve Bayes is acknowledged as graphical modeling tool and used to model decision problems having uncertainty. Naïve

Bayes offer better detection rate on three main attacks such as probing, user to root and remote to user. Decision Tree have better accuracy [7]. K-Nearest Neighbor is also robust to noisy data by averaging k-nearest neighbors. By combining the three algorithms this research can achieve the best performance in detecting intrusions.

```
Input: Test Dataset
Process:
BEGIN
Read test dataset
Call Naïve Bayes and Decision Tree to classify as
anomaly or normal
If status is anomaly then
Sent report
End If
END
Output: Report
```

Pseudo code 1: Pseudo code for training phase

4. Inference Engine

After the Naïve Bayes, Decision tree and K-Nearest Neighbor model are built or trained by the network traffic dataset and ready to predict attacks in the incoming network traffic, the Inference Engine provides the predicting algorithms with test data which is used to compare with the learnt knowledge. The Inference Engine also classify each record of the input data (Test dataset) to normal connection or to a relevant attack based on the combined model. The integration of the algorithms are happened after the three algorithms made prediction. Then when they made same prediction, the integrator code takes the result but when they predict different the integrator code takes the worst possibility which is the anomaly.

```
Input: New Dataset with reduced dimension
Process:
BEGIN
Try
Read features from dataset with reduced dimension
Train Decision tree classifier using training dataset
Train Naïve Bayes classifier using training dataset
Train ibk-KNN classifier using training dataset
End Try
END
Output: The trained model
```

Pseudo code 2: Pseudo code for detection stage

3.3.Execution Flow of the Proposed System

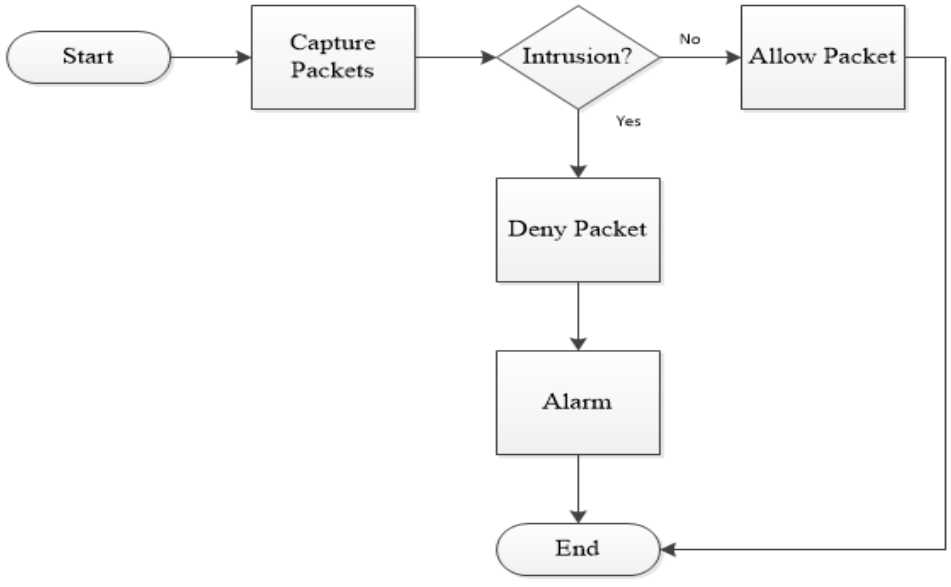

Figure 2: Execution flow of the proposed system

4. IMPLEMENTATION AND EVALUATION

The proposed system is developed and implemented for network based attacks. The WEKA API is also used for data mining and machine learning purpose in developing the intrusion detection system.

4.1.Experiment on Network based Intrusion Detection System

The intrusion detection system uses combined algorithms such as Naïve Bayes, Decision tree and KNN. Since the main aim is to reduce the false alarm rate, the results in contrast with other algorithms will be discussed on next pages. The

next figure shows how the proposed combined system trained with the training dataset with reduced dimension and then detect intrusive actions from the new dataset.

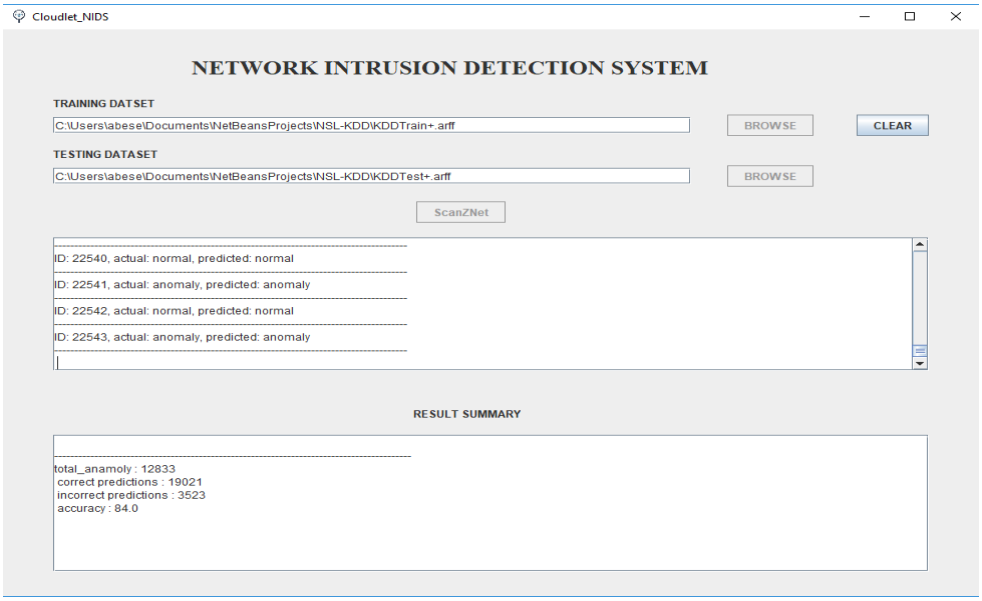

Figure 3: Performance of the combined algorithm

4.2.Performance Evaluation

The performance of the proposed network based intrusion detection system is evaluated based on the accuracy, precision, recall, True Positive Rate (TPR) and False Positive Rate (FPR). To measure the performance of the NIDS a standard metrics which is confusion matrix values are used.

The effectiveness of network based intrusion detection system is measured in terms of accuracy in which it identifies how much do the IDS classify the incoming packet as normal and attack. The accuracy of the proposed system is calculated using the following equation.

$$\text{Accuracy} = \frac{(\text{correctly predicted} * 100)}{(\text{correctly predicted} + \text{incorrectly predicted})}$$

This research first calculate the accuracy of each algorithm which is Naïve Bayes, Decision Tree and ibk-KNN. After that the research compares the result with the combined algorithm. For the algorithms the research uses 22544 instances. The experiment got 17160 correctly predicted and 5384 incorrectly predicted instances for Naïve Bayes algorithm, 17913 correctly predicted and 4631 incorrectly predicted instances for Decision Tree algorithm and 17890 correctly predicted and 4654 incorrectly predicted instances for ibk-KNN algorithm. After calculating the accuracy of each algorithm the result for Naïve Bayes is 76%, for Decision Tree 79% and for ibk-KNN 79%. But in contrast to the results from the Naïve Bayes, Decision Tree and ibk-KNN algorithms, the combined Naïve Bayes and Decision Tree algorithm produces 18543 correctly prediction and 4001 incorrectly prediction. From this result the accuracy of combined NB and DT algorithm is 82%.

Conversely, the combined NB, DT and ibk-KNN produces 19140 correctly prediction and 12833 incorrectly prediction. From this result the accuracy of the combined NB, DT and ibk-KNN algorithm is 84%. Therefore the accuracy of the combined NB, DT and ibk-KNN algorithm is better than that of the single algorithm and the combined NB and DT.

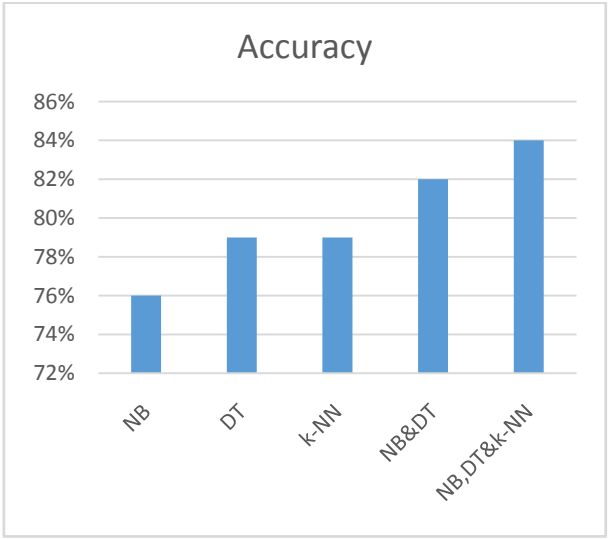

Figure 4: Accuracy of the algorithms

The next figures shows the performance of both the Naïve Bayes, Decision tree and ibk-KNN algorithms giving the same datasets.

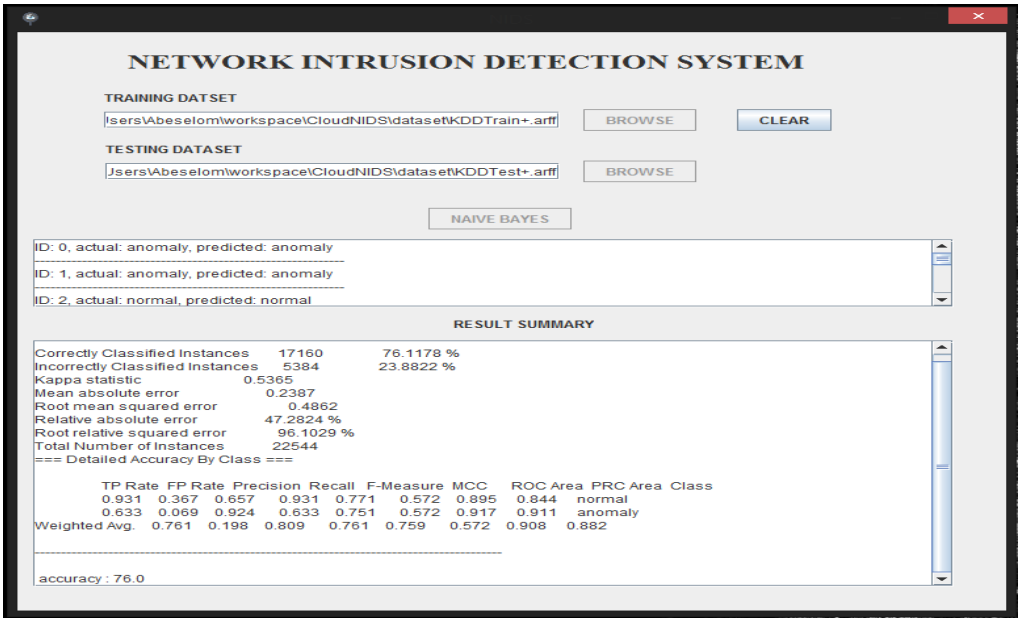

Figure 5: Performance of Naïve Bayes algorithm

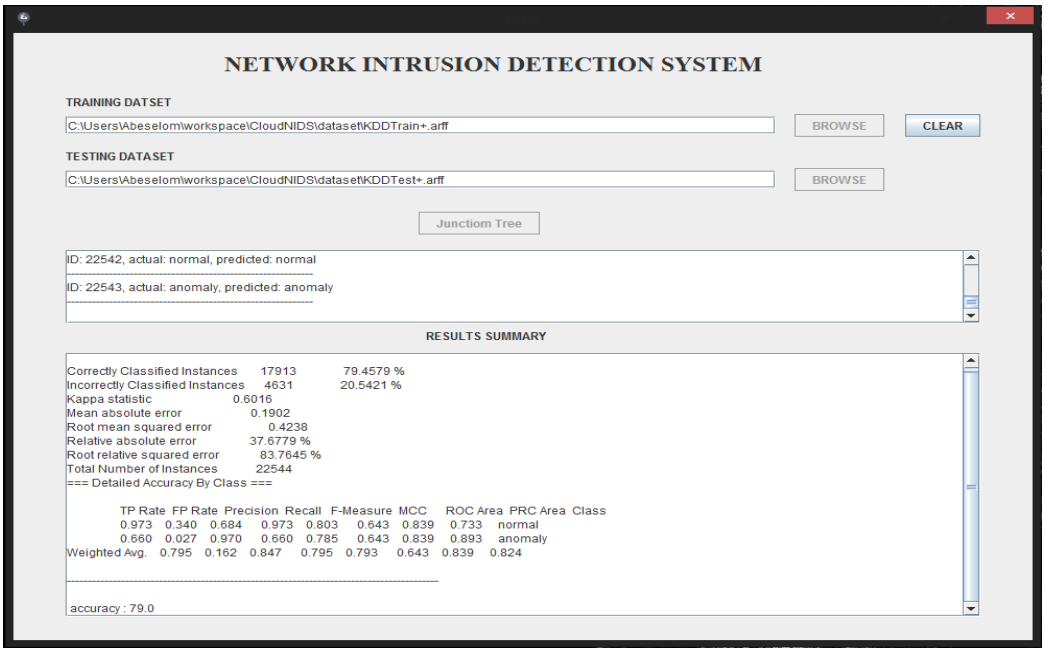

Figure 6: Performance of Decision Tree algorithm

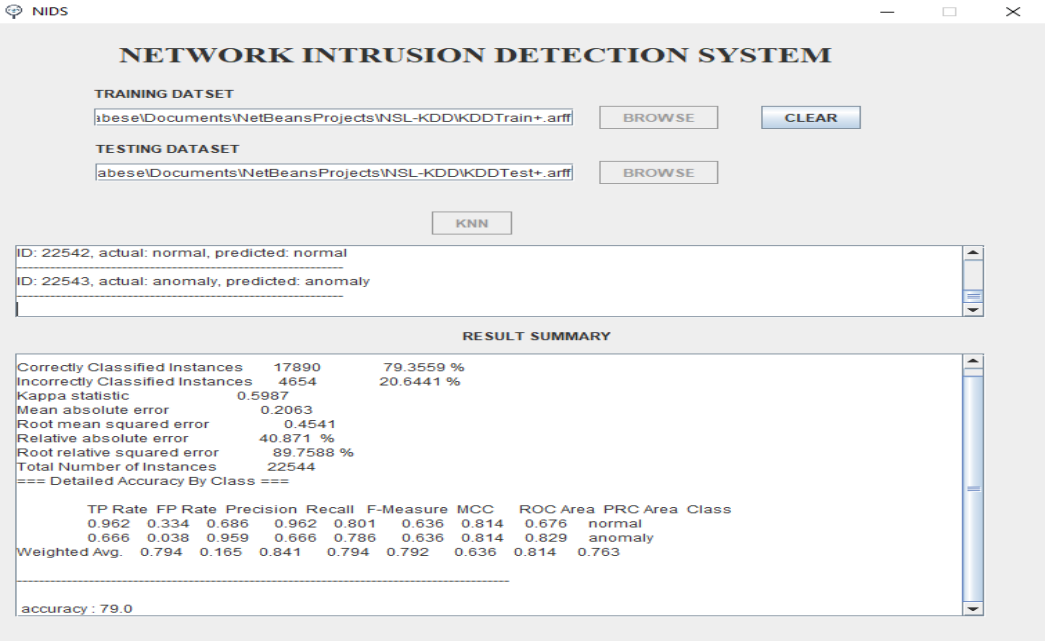

Figure 7: Performance of ibk-KNN

5. CONCLUSION AND FUTURWORK

5.1.Conclusion

Network intrusion detection in hostile network environment has been improved from time to time. However, this improvements have not been enough to secure the network users from cyber-attack. This research tried to fill the gap around the short coming of the current network based intrusion detection system by proposing combined network intrusion detection system. The proposed system intended to deter data theft by intruders or reduce the vulnerability of the network.

The proposed system used the combined algorithms which are Naïve Bayes, Decision Tree and K Nearest Neighbor to classify the incoming network traffic. The anomaly based component is trained using simulated dataset called NSL-KDD.

Finally, the proposed combined network based intrusion detection system is evaluated and the obtained result showed the system can deter intrusions. This showed that the proposed system can be effective means of security for computer networks.

5.2.Future work

Beside the promising result of the proposed network based intrusion detection system, there is a need to improve some aspects of the system. This research believed that more research is needed to make the system more effective and accurate.

Some potential future work that could be a continuation of this research work is as follows:

- Developing a more accurate model that can be used in real-time for detecting and classifying anomaly with minimum false alarms and less time.
- Implementation of the proposed network intrusion detection system in real time network environment.

6. ACKNOWLEDGMENTS

First and foremost, I would like to thank my almighty God for making this possible. My deepest gratitude also goes to my family and friends for giving me invaluable hope, courage, ideas, and comments.

7. REFERENCES

[1] C. A. K. Sumaiya Thaseen, "analysis of supervised Tree based classifiers for intrusion detection system," International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013.

[2] R. K. C. Nidhi Srivastav, "Novel Intrusion Detection System integrating Layered Framework with Neural Network," IEEE, 2012.

[3] M. P. a. M. R. Patra, "A Comparative Study Of Data Mining Algorithms For," IEEE, pp. 504-507, 2008.

[4] Q. Y. D. R. Juan Wang, "An intrusion detection algorithm based on decision tree technology," IEEE, 2009.

[5] Y. Freund, "The Alternating Decision Tree Algorithm," ICML, pp. 124-133.

[6] D. Tigabu, "Constructing Predictive Model for Network Intrusion Detection," 2012.

[7] S. B. a. Z. E. N. Ben Amor, "Naive Bayesian Networks in Intrusion Detection Systems," 2000.