Analysis of Ensemble and Hybrid Approaches for Intrusion Detection

Abhinav Jaiswal (2016IPG-004)
Mohit Kumar (2016IPG-053)
Surendra Singh Gangwar(2016IPG-107)

ABV-IIITM Gwalior
Gwalior, MP, India

Introduction

- In recent years, there has been a revolutionary change in the field of networking, use of internet is growing day by day and so is the risk of intrusion, Nowadays, it is very important to have a secure network because of increasing dependability on the internet and to achieve a high level security.
- One of the possible ways to counter this problem is through intrusion detection, which aims to identify various network attacks.
- Advancement in machine learning and deep learning made intrusion detection algorithms result in less error rate and more accurate to classify in less possible time.

Motivation

- The methods used for intrusion detection have proven to be advantageous but classification of different intrusion attack type efficiently is still a major concern
- Data-set used in the past was inconsistent,hence their is a need of consistent data-set to classify the attack accurately.
- Our main work is to combine the ability of different machine learning algorithm to accurately detect a particular attack type, and use them to build a model which is cable of predicting all attack type precisely. .

Literature Review

- Many different types of machine learning approaches has been implemented by researchers to classify the intrusion, some has used single learning algorithms[2] such as Support Vector Machine(SVM)[3] ,Logistic Regression(LR) , K-nearest Neighbors(KNN) and some uses multilevel techniques in which they first use some nature inspired algorithm[1] and genetic algorithm to do feature selection and then used single classifier to predict the result.
- The challenges faced in Intrusion detection has been studied by many researchers around the globe .They suggested that their is need of efficient methodology which can identify any kind of intrusion attack precisely with the goal that a specific counter measure could be taken.

- Problem Statement
  - Analysis of ensemble and Hybrid Approaches for Intrusion Detection.
- Thesis Objective
  - Evaluating the performance of many different machine learning algorithms and ensemble approaches for the intrusion detection.
  - Implementation of mixture of experts technique for intrusion detection.
  - To perform a comparative analysis of ensemble approaches(Bagging,Boosting)with hybrid approach(Mixture of experts).

Methodology Used

- Observation of the KDD 1999 Dataset.
- Data cleaning and preprocessing.
- Division of entire dataset into five input space labeled with Normal, Dos, Probe, U2R, R2L.
- Evaluation of the performance of different base learners with the given dataset.
- Implementation of the Bagging, Boosting and Mixture of experts with the selected base learners.
- Comparative analysis of ensemble and hybrid approaches on the basis of accuracy score, precision, recall and F1-score.

Ensemble Approaches

- Bagging: It comprises of two approaches, first it randomly chooses bootstrapped samples from the given dataset and build a classifier for each bootstrap sample and then aggregate the reults from all classifiers.

- Boosting: Boosting is a sequential learning algorithm which trains weak learners to converts them into the strong learner. In each iteration, it tries to increase the weights of poorly predicted instances.

Work Flow

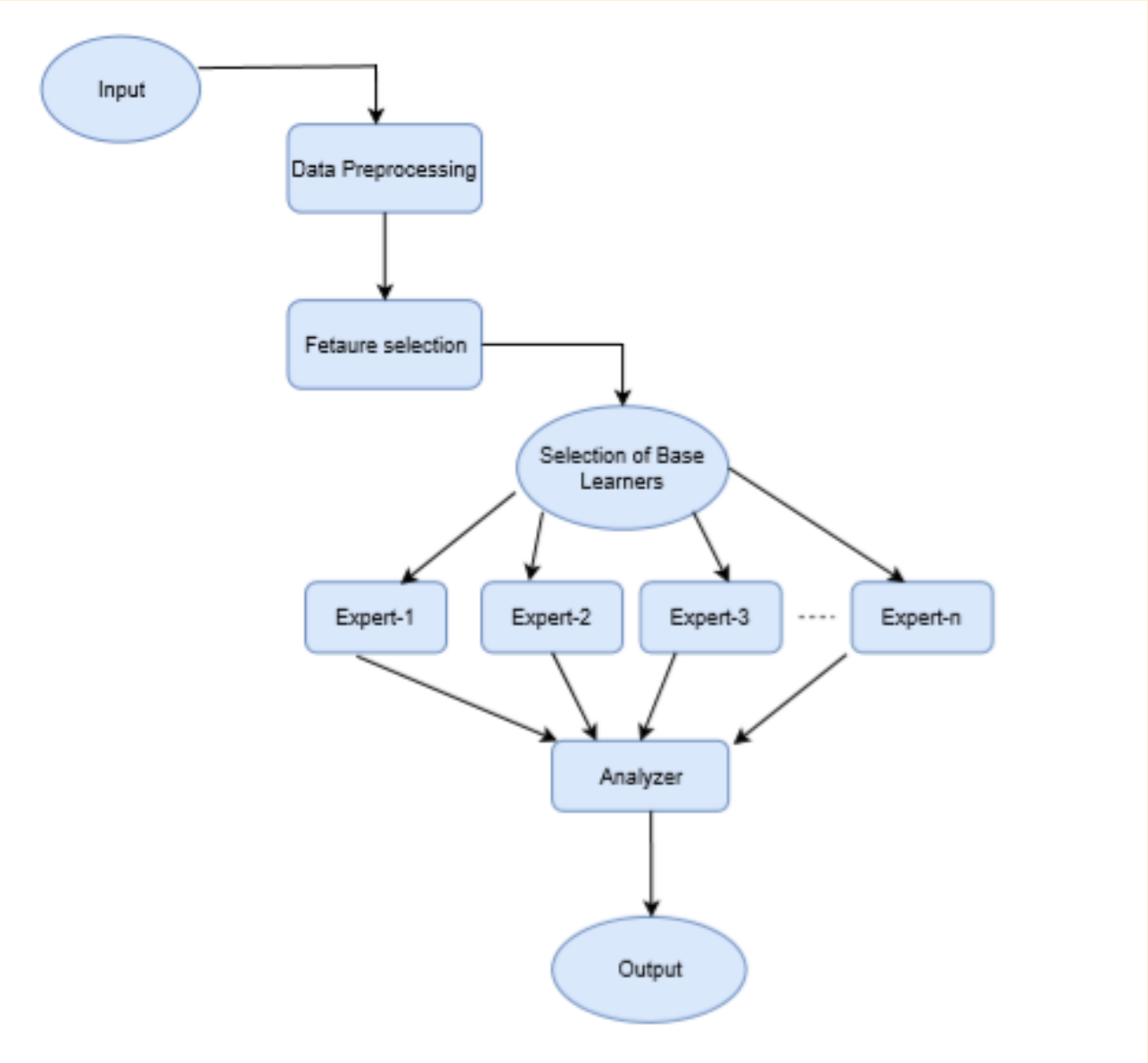


Figure: 1 Work Flow

Architecture of MoE

Input to the architecture comprises of data with selected features using feature selection algorithm on the dataset,working of the model can be explained in the following steps:

- first all the selected experts predict the attack type for the corresponding tuple and after that to get the required result voting for the particular attack type is done.
- After voting if their exist a class which clearly outweight all the class than that class is the required result if the probability of all the class to be of that particular attack type is equally probable than the output is taken from the experts which are expert in the particular attack type.
- All the predicted output from the experts are given the dynamic weight on the basis of that output whose probability is highest is accepted as the final output for the given tuple.

Architecture of MoE

The aim of the model is to correctly classify all the classes either by using the voting method or by using epxertise of different experts to get the desired output using the above approach the probability of getting a wrong output minimizes and we get the high accuracy result.

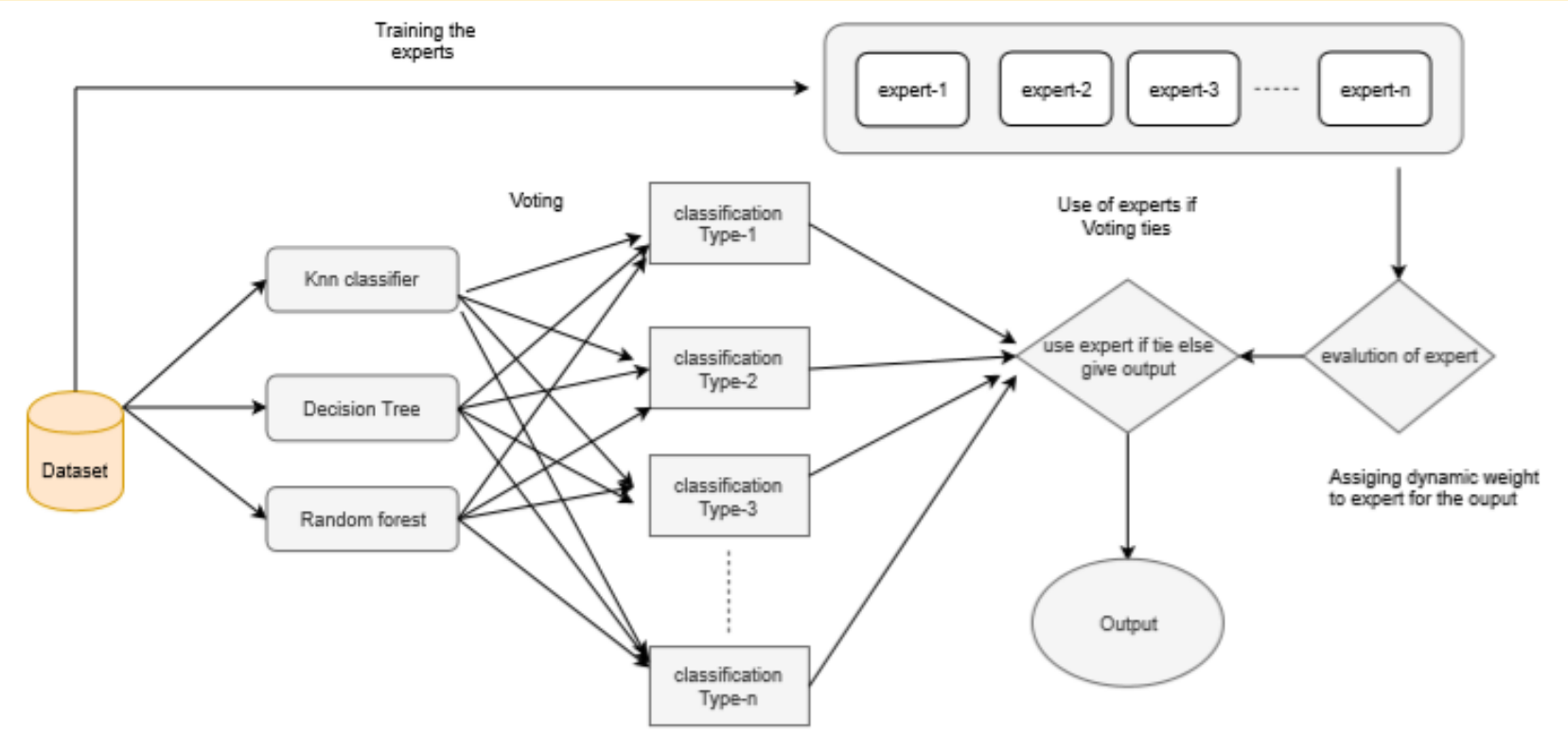


Figure: 2 Work Flow

Results and Analysis

Classifer	Accuracy	Precision	recall	F1-score
LinearSVC	0.9221	0.9229	0.9147	0.9187
KNN	0.9690	0.9699	0.9647	0.9674
MultinomialNB	0.4172	0.4161	0.4159	0.3966
RandomForest	0.9688	0.9679	0.9610	0.9645
LogisticRegression	0.8959	0.8912	0.8899	0.8457
DecisionTree	0.9490	0.9474	0.9447	0.9447

Figure: 3 Intermediate Results

Base classifiers Results

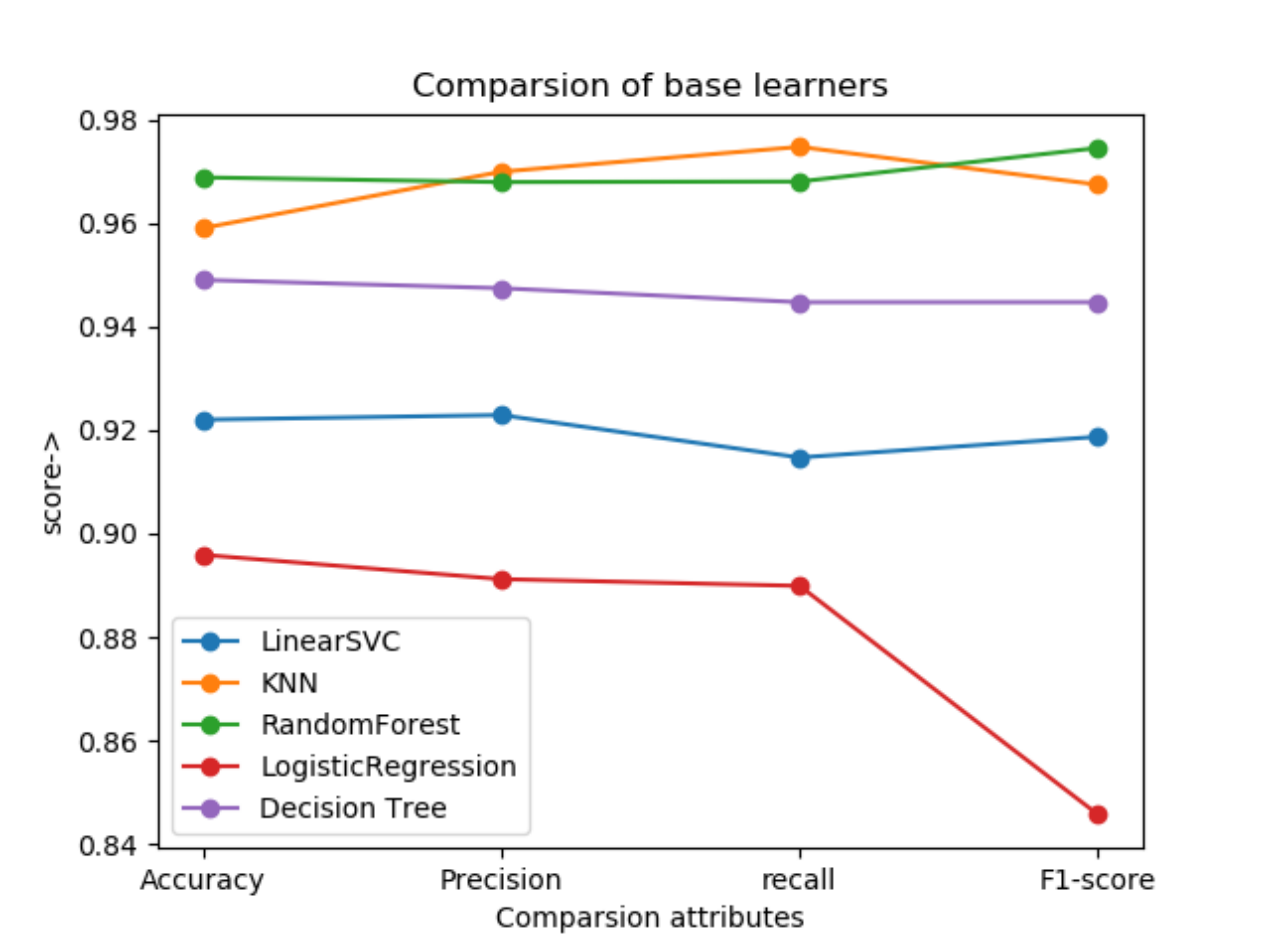
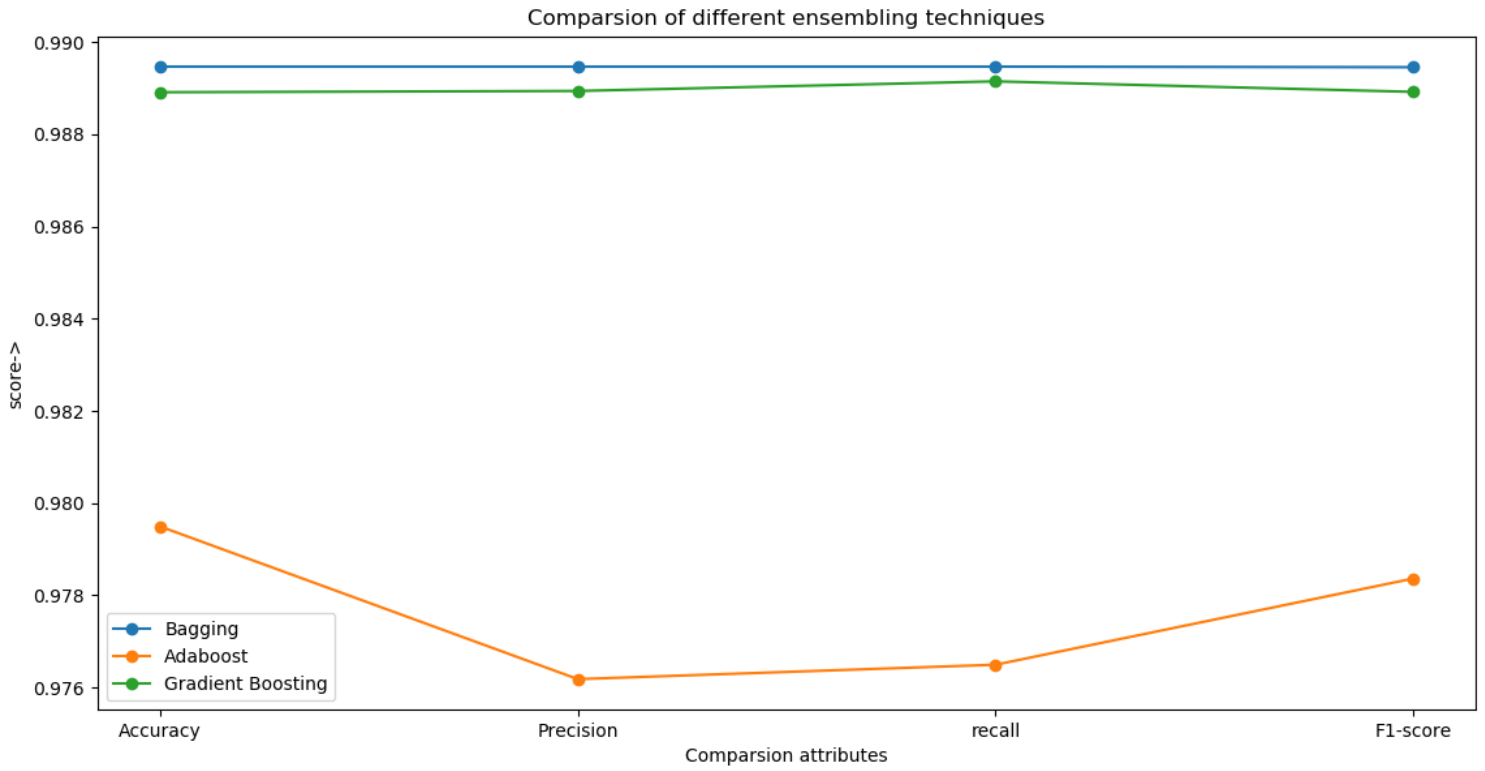


Figure: 4 Base Classifiers Comparison

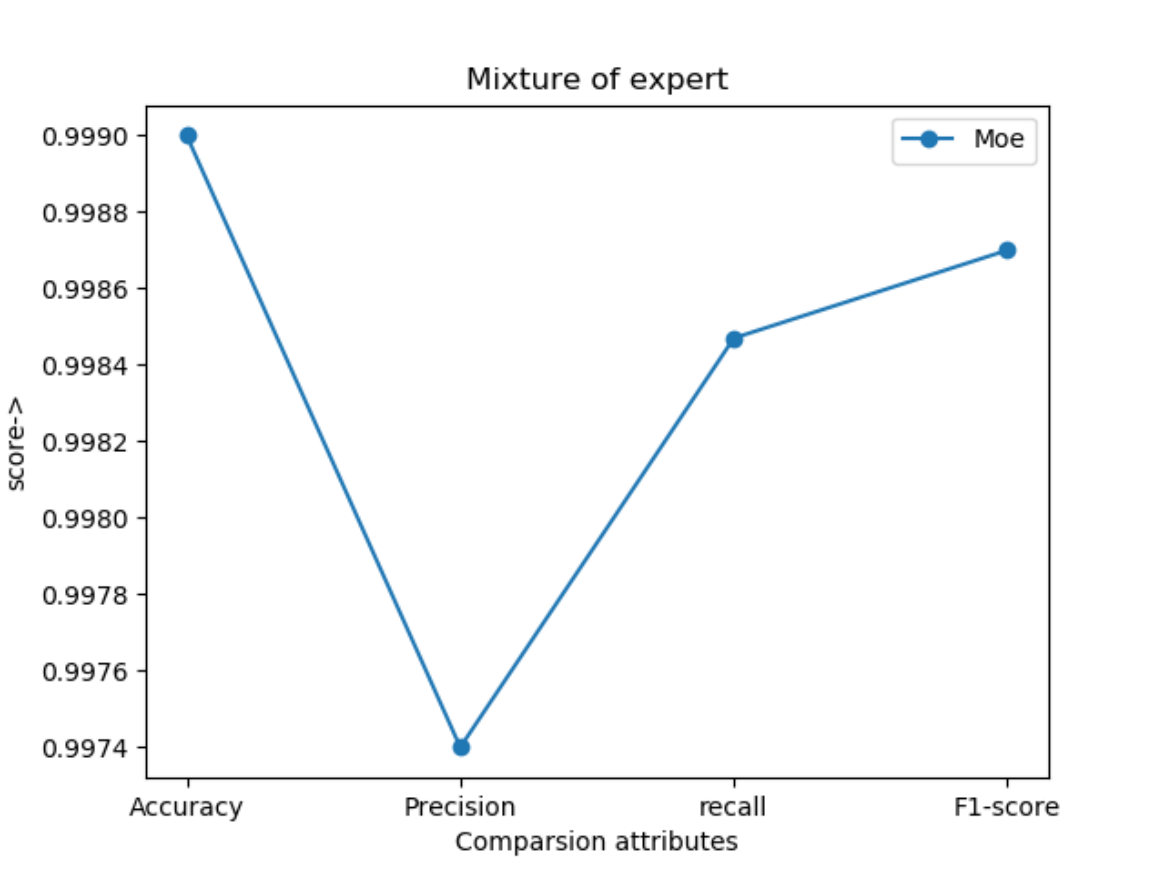
Ensemble: Bagging and Boosting

Classifier	Accuracy	Precision	recall	F1-score
Bagging	0.98947	0.98947	0.99847	0.98945
Adaptive Boosting	0.97954	0.95218	0.97654	0.97836
Gradient Boosting	0.98891	0.98894	0.98891	0.98892



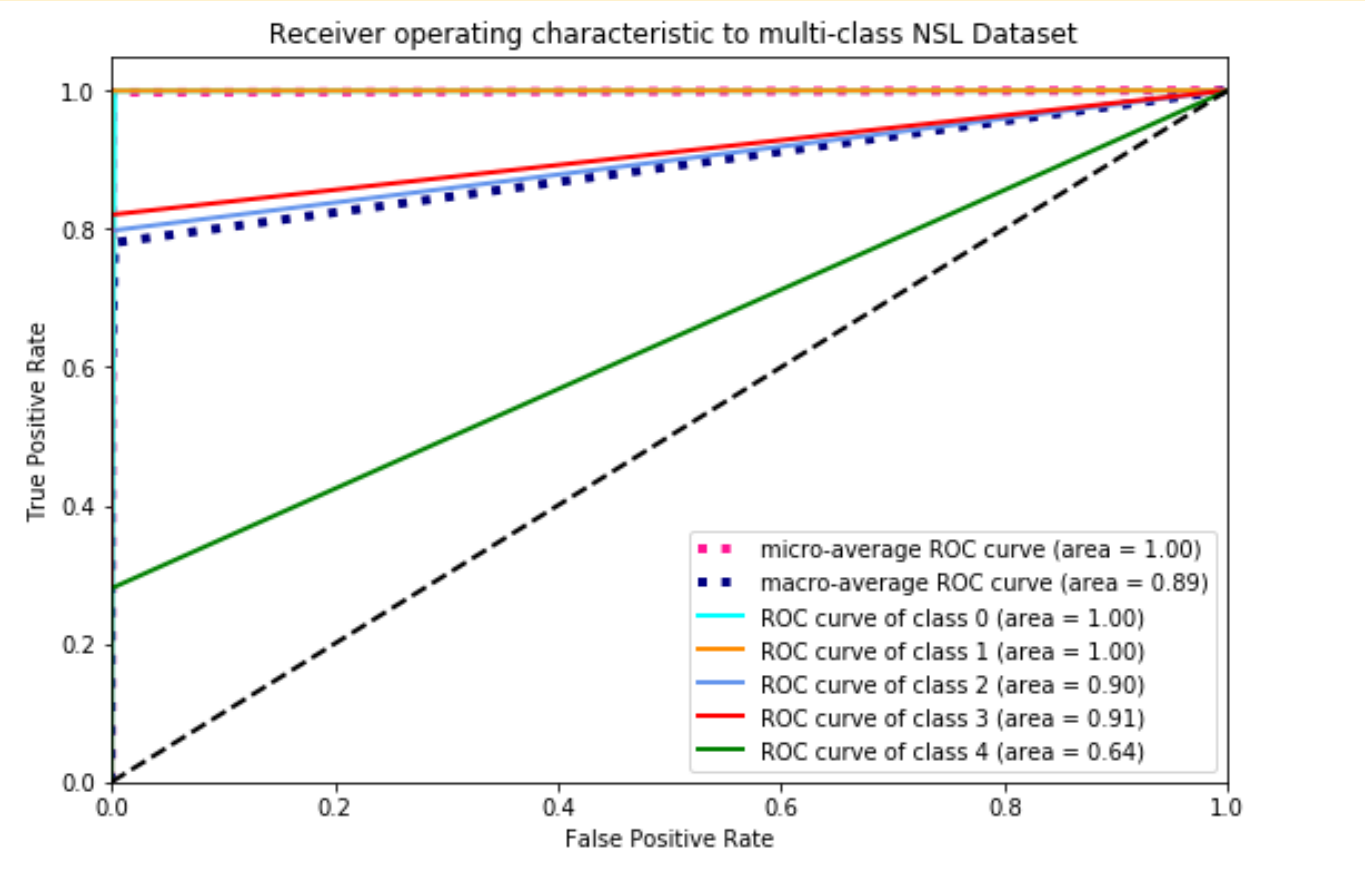
Proposed model result(MoE)

Classifier	Accuracy	Precision	recall	F1-score
MoE	0.99901	0.99745	0.99847	0.99987



AUC-ROC for MoE

Auc-roc is used to measure the performance for the classification at various thresholds settings , ROC represents the probability curve and AUC speaks to degree or proportion of separability.By similarity, Higher the AUC, better the model is at recognizing different attack type



Web Application

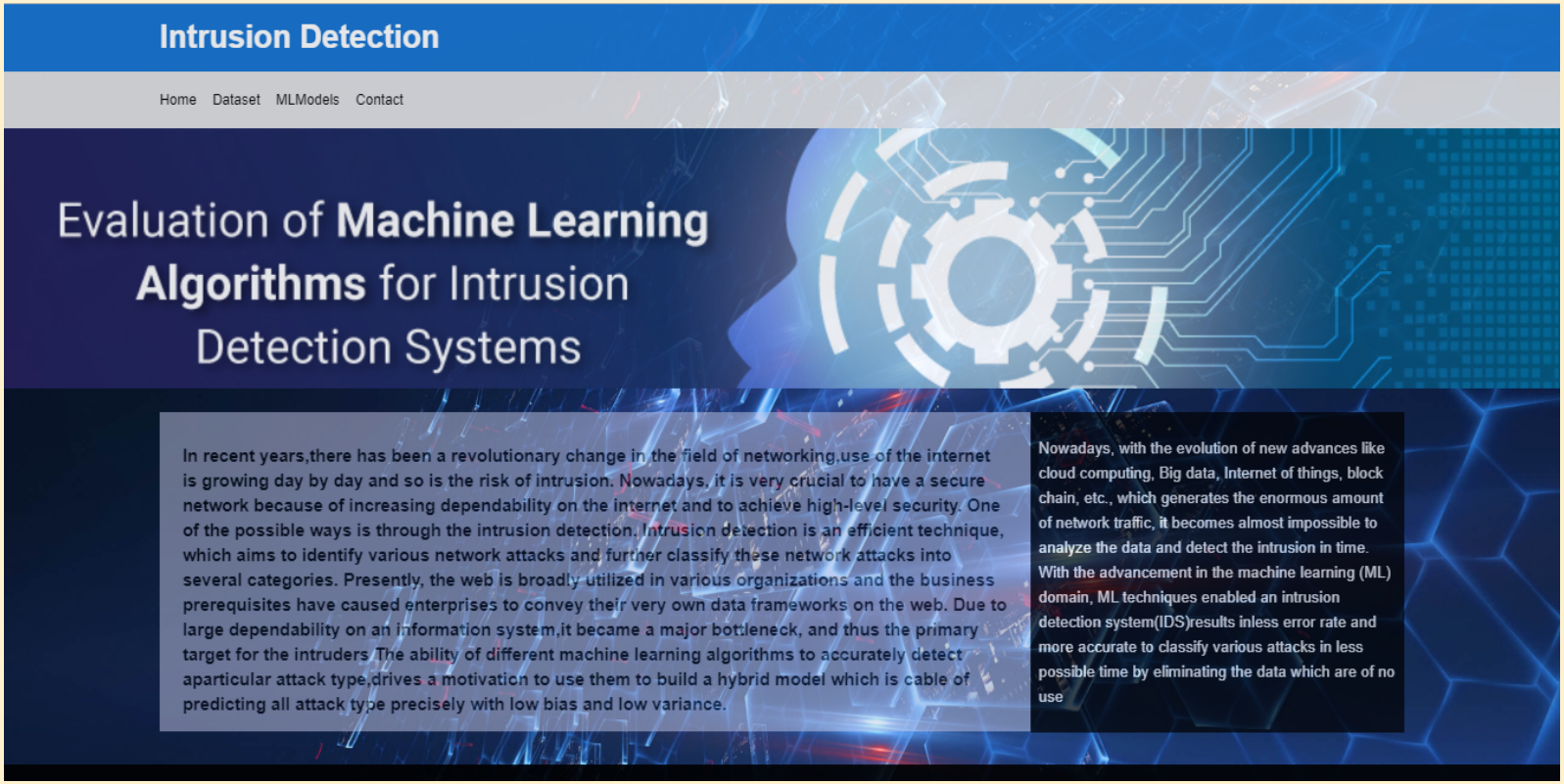


Figure: 7. Homepage

Contd...

NSL - KDD 1999 Dataset

The KDD 1999 dataset was developed by the MIT Lincoln Labs and was extensively used by researchers during the last decade. The entire dataset is very large in size and contains many attributes variables. Therefore to improve the machine learning computation, 10 % of it was extracted and adopted as training dataset in the intrusion detection process. However, some inherent drawback was made about this dataset. The KDD 99 contains an important quantities of redundant records which has as consequence to prevent the learning algorithm to perform well.To resolve some issues found in the previous KDD 99, an improved version was created, the NSL-KDD dataset.The classes or labels in the NSL KDD dataset are divided into four categories which represent the attack class and one as normal traffic.

Four categories of Attacks:

1. Denial of Service (DoS): This attack aims to block or restrict a computer system or network resources or services.
2. Probe: The intruder aims to scan for information or vulnerabilities in a network or computer system which later on will be used to launch attacks.
3. Remote to Local (R2L): Here the intruder gain remotely unauthorized access to a computer system over a network by sending data packet to that system.
4. User to Root (U2R): Here the intruder gains access to a user with normal privilege and later on try to access a user with administrator or root privilege.

The reason behind the use of this dataset are following relevant to mention:

- Elimination of redundant records in the training set will help our classifier to be unbiased towards more frequent records.
- No presence of duplicate records in the test set, therefore, the classifier performance will not be biased by the techniques which have better detection rates on the frequent records.
- The training and test set contains both a reasonable numbers of instances which is affordable for experiments on the entire set without the need to randomly choose a small portion.

Figure: 8. Dataset

Contd...

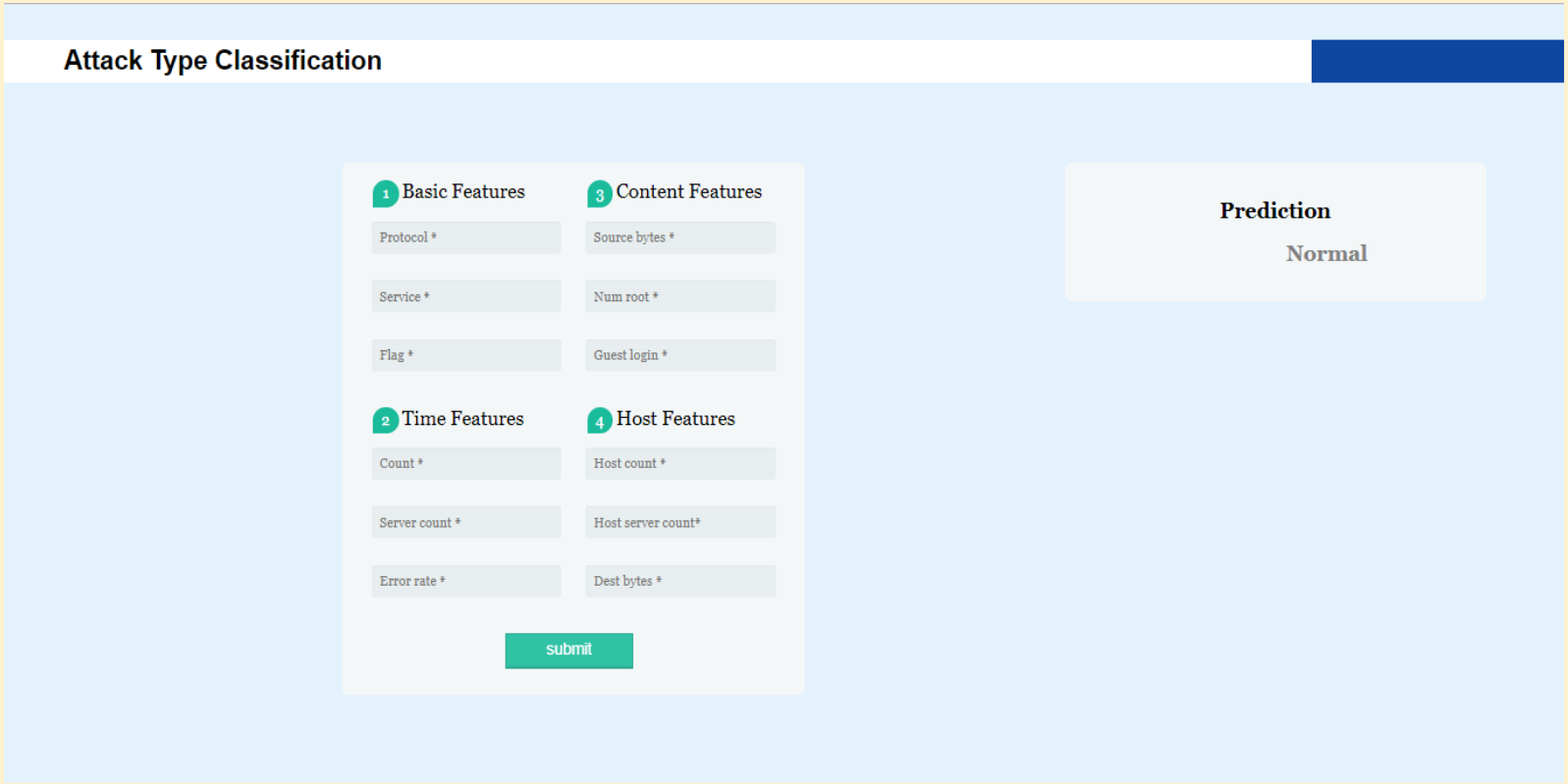


Figure: 9. Normal traffic Flow

Contd...

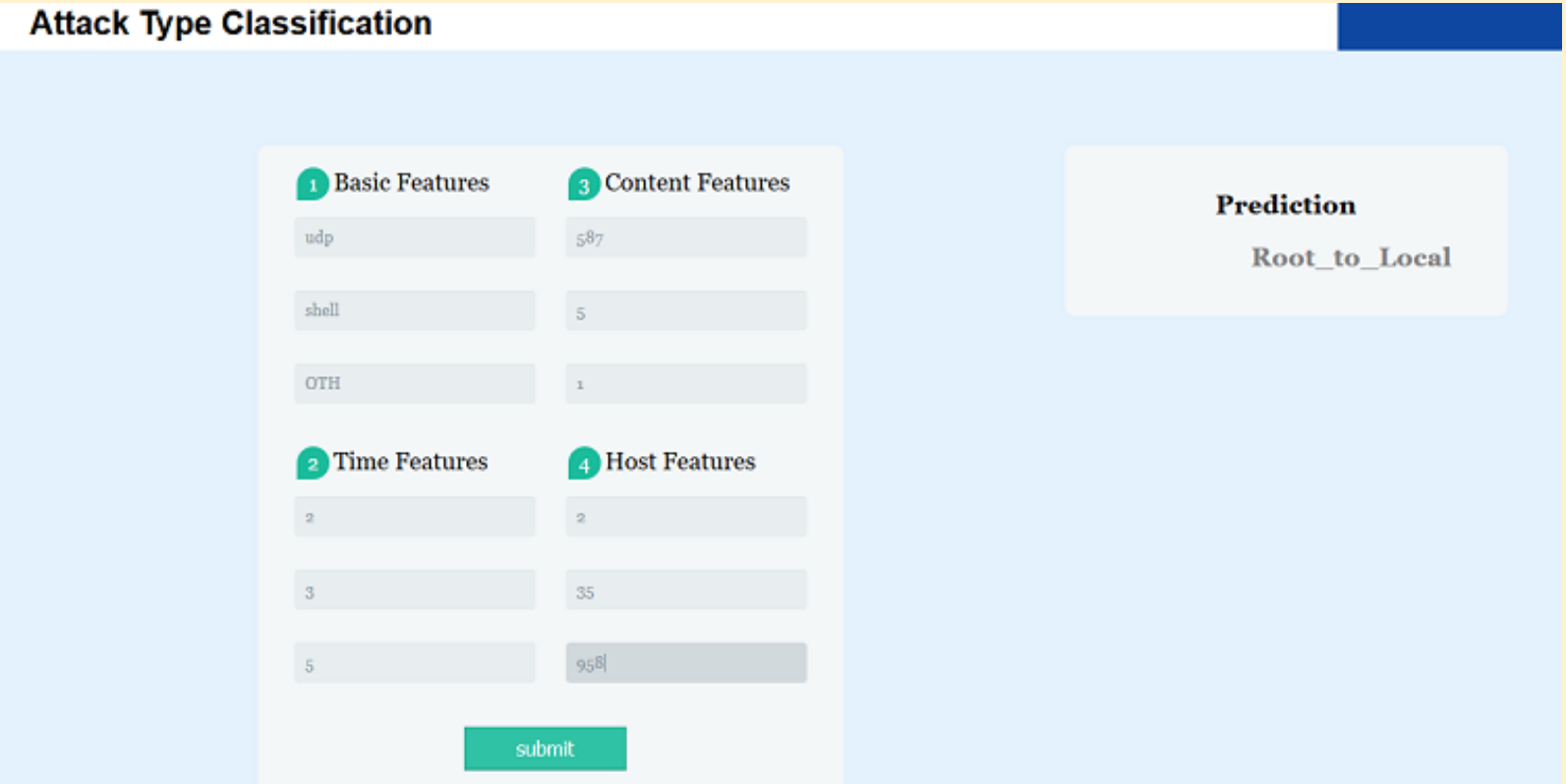


Figure: 10. Malicious Traffic Flow

Future Activities

- Implementation on Large dataset
  - Implementation of the proposed model with the large dataset, rich in wide variety of attacks.
- Implementation on other areas
  - Applications of mixture of experts technique has certainly exhibit tremendous results. Therefore,try to implementing this technique on other areas.
- Dynamic web interface
  - Building a web based application to detect the type of attack generated dynamically on the system.

REFERENCES

- B.Selvakumar and K.Muneeswaran , "Firefly algorithm based feature selection for network intrusion detection", *Computer and Security*, vol. 81, pp. 148-155, 2018.
- C.F. Tsai, Y.F. Hsu,C.Y Lin, W.Y. Lin, "Intrusion detection by machine learning: A review", *Expert Systems with Applications*, vol. 10, pp. 11994-12000, 2009.
- Gu, Jie and Wang, Lihong and Wang, Huiwen and Wang, Shanshan,"A novel approach to intrusion detection using SVM ensemble with feature augmentation", *Computers Security*, vol. 1, pp. 1, 2019.

Thank You