International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
www.computersciencejournals.com/ijcit
IJCIT 2021; 2(1): 05-09
Received: 09-11-2020
Accepted: 19-12-2020

Manish Khule
Assistant Professor,
Department of CSE, Symbiosis
University of Applied Sciences,
Indore, Madhya Pradesh,
India

Neha Sharma
Assistant Professor,
Department of CSE, Symbiosis
University of Applied Sciences,
Indore, Madhya Pradesh,
India

Corresponding Author:
Manish Khule
Assistant Professor,
Department of CSE, Symbiosis
University of Applied Sciences,
Indore, Madhya Pradesh,
India

Anomaly detection model based on SVM & XGBoost to detect network intrusions

Manish Khule and Neha Sharma

DOI: https://doi.org/10.33545/2707661X.2021.v2.i1a.21

Abstract

There are rapidly increasing attacks on computers creates a problem for network administration to prevent the computer from these attacks. There are many traditional intrusion detection systems (IDS) is present but they are unable to prevent computer system completely. The need to secure networks has increased as the number of people connecting to the network are increasing rapidly and using networks for storing or accessing critical information. In this paper we have assessed and compared various machine learning algorithm and then propose a system based on the best performing algorithm. In this we proposed an XGBoost learning technique which is an combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model.

Keywords: Network security, intrusion detection system, data mining algorithm, machine learning techniques, anomaly detection, SVM, XGBoost

1. Introduction

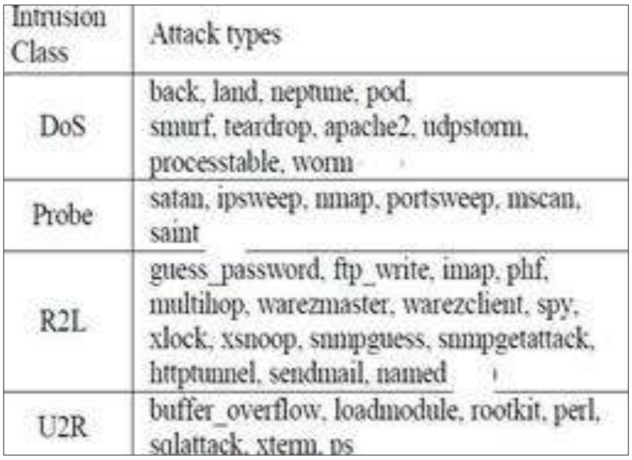
With the advancement in the technology, millions of people are now connected with each other through one or other form of network where they share lots of important data. Hence the need of security to safeguard data integrity and confidentiality is increased rapidly. Although effort have been made to secure data transmission but at the same time, attack technique for breaching the network continued to evolve. Thus it leads to the need of such a system which can adapt with this ever changing attack techniques. In this paper, we have purposed a system which is based on machine learning. Our aim is to find the based suitable machine learning algorithm which can predict the type of network attack with highest accuracy and then develop a system which uses this algorithm to detect network intrusion. The algorithms which we have compared are SVM and XGBoost model. The dataset used for training the model is KDD 99 dataset. The reason why we have used machine learning is the flexibility that it provides to the system for example, if any new type of attack is developed in future the system can be trained for predicting that attack. There are a few types of intrusion detection system out of which ours is a knowledge based intrusion detection system which is also known as the anomaly based system.

It registers the anomalies and in future predicts such malicious network to send out an alert. This way the network can disconnect to the such a connection and then have only secured connections.

1.1 NSL-KDD Dataset

The data used in this study is the NSL-KDD Cup 1999 dataset. The NSL-KDD is the proposed dataset in several studies as a solution to the problem in the KDD Cup 1999 dataset (KDD-99). The KDD-99 dataset is over 15 years old [5], but is still commonly used in research on the area of intrusion detection systems due to the lack of datasets available to the public and accessible freely. Some of the problems that exist in the KDD-99 dataset have now been addressed in NSL-KDD including deletion of redundant data and re-proportion of datasets [4]. NSL- KDD does not include redundant data on KDD-99 which may affect the performance of learning algorithms while the re-proportion of KDD-99 enables NSL-KDD in the process of evaluating various learning algorithms. The NSL-KDD dataset has 42 attributes, which consists of 41 input attributes and 1 target attribute. Furthermore the types of attacks are grouped into four categories of intrusion classes namely DoS, Probe, U2R and R2L. Table I shows classes and members (attack type) for eachclass.

Table 1: Attack Class



Intrusion Class	Attack types
DoS	back, land, neptune, pod, smurf, teardrop, apache2, udpstorm, processtable, worm
Probe	satan, ipsweep, nmap, portsweep, mscan, saint
R2L	guess_password, ftp_write, imap, phf, multihop, warezmaster, warezclient, spy, xlock, xsnoop, snmpguess, snmpgetattack, httptunnel, sendmail, named
U2R	buffer_overflow, loadmodule, rootkit, perl, sqlattack, xterm, ps

2. Literature Review

According to [1], software package outlined networking could be a construct projected to interchange ancient networks by separating management plane and information plane. It makes the network additional programmable and manageable. As there's one purpose of management of the network, it's additional liable to intrusion. The thought is to coach the network controller by machine learning algorithms to let it create the intelligent choices mechanically. During this paper, we've mentioned our approach to form software package outlined networking safer from varied malicious attacks by creating it capable of police work and preventing such attacks.

In [2] IDS area unit outlined as a software package utility that detects system activities for risky movements and generates experiences to management. Associate degree Intrusion Detection System is security counter step to acknowledge set of intrusion that compromises the acquaintance, availableness, and integrity of information sources [1]. The teams area unit victimisation IDS with aim to spot issues with security policies and documenting existingthreats.

In [3] we have a tendency to gift the results of our experiments to judge the performance of police work differing types of attacks (e.g., IDS, Malware, and Shellcode). We have a tendency to analyze the popularity performance by applying the Random Forest rule to the varied datasets that area unit made from the Kyoto 2006+ dataset, that is that the latest network packet information collected for developing Intrusion Detection Systems. We have a tendency to conclude with discussions and future analysiscomes.

3. Problem Definition

With the rapid development of information technology in the past two decades. Computer networks are widely used by industry, business and various fields of the human life. Therefore, building reliable networks is a very important task for IT administrators. On the other hand, the rapid development of information technology produced several challenges to build reliable networks which is a very difficult task. There are many types of attacks threatening the availability, integrity and confidentiality of computer networks. The Denial of service attack (DOS), probe, R2L and U2R are considered as of the most common harmful attacks.

4. Proposed Work

In this paper, we proposed an network IDS based on XGBoost algorithm classifier. These classifier enhances the attack detection accuracy and it very efficient in distinguishing network traffic is attack or normal.

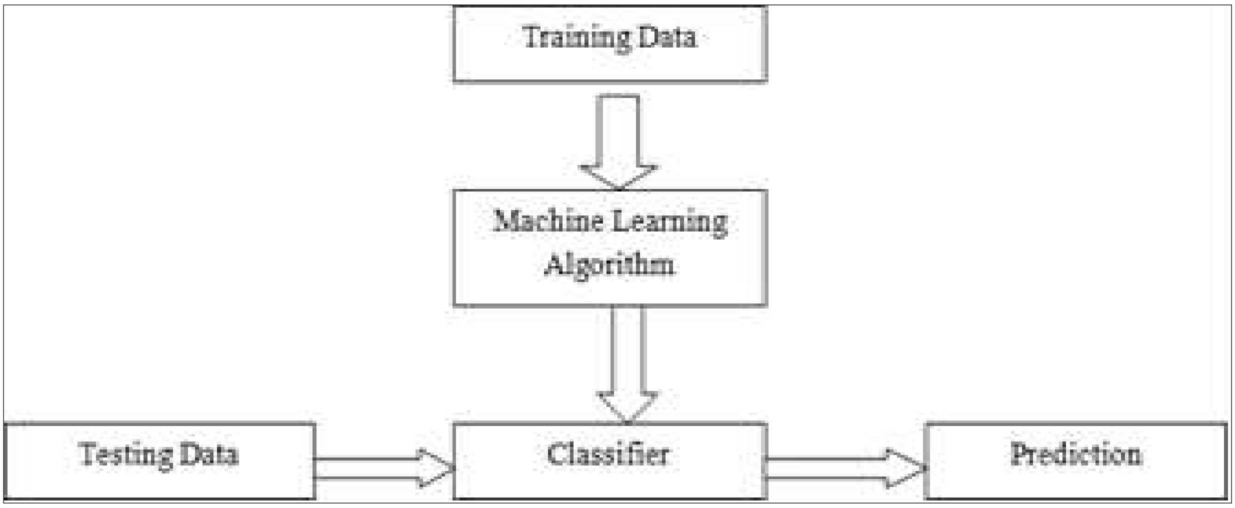


Fig 1: Proposed Model

Machine Learning

One of the most challenges for IDSs is to make effective behaviour models or patterns to tell apart traditional behaviours from abnormal behaviours by observant collected audit information to resolve this drawback, earlier IDSs sometimes place confidence in security consultants to analyse the audit information and construct intrusion detection rules manually. Since the quantity of audit information, will increase vary quick, it's become a long, tedious and even not possible work for human consultants to analyse and extract attack signatures or detection rules from dynamic, immense volumes of audit information. conjointly the detection rules created by human consultants area unit sometimes supported fastened options or signatures of existing attacks, therefore it'll be terribly tough for these rules to observe malformed or perhaps utterly newattacks.

Due to the on top of deficiencies of IDSs supported human consultants, intrusion detection techniques victimisation data processing have attracted a lot of and a lot of interests in recent years. As a very important application space of knowledge mining, intrusion detection supported data processing algorithms, that is typically said as reconciling intrusion detection, aims to resolve the issues of analysing immense volumes of audit information and realizing performance improvement of detection rules.

By creating use of knowledge mining algorithms, reconciling intrusion detection models are often mechanically created supported labelled or untagged audit information.

A methodology for intrusion detection is projected that involves a attribute choice method for choice of relevant attributes and there on applying a classifier for classifying network information to 2 categories: normal categories and attack categories.

5. Experiment Alanalysis

The dataset taken from the Kdd99 is a huge dataset and the one that we have used in our research. Our aim is to not only to find the best algorithm suited for the intrusion detection but also to implement it using the programming language R. In this we proposed an XGBoost learning algorithm for detecting intrusion. The first process of applying learning is to pre-process the data. For which we prepare an R scripts to add the names of the columns and to group the data of the attacks according to five classes (DoS, Probe, U2R, R2l, normal) [5].

Table 2: Number of Samples

Type	Number of samples in training set	Number of samples in test set
DOS	3514	1486
Normal	5232	2268
Probe	2090	886
R2L	261	118
U2R	36	14
Total	11133	4772

Then we have to remove the redundant rows from the dataset. Then our next step is to see whether there are any missing values and then to remove those corresponding rows too. After the dataset preparation and pre-processing we can do exploratory data analysis in which we can explore the data by visualizing to better understand the data. Figure 2 shows the protocol which having high frequency of attacks, and figure 3 shows the correlation between duration, attacks and protocoltypes.

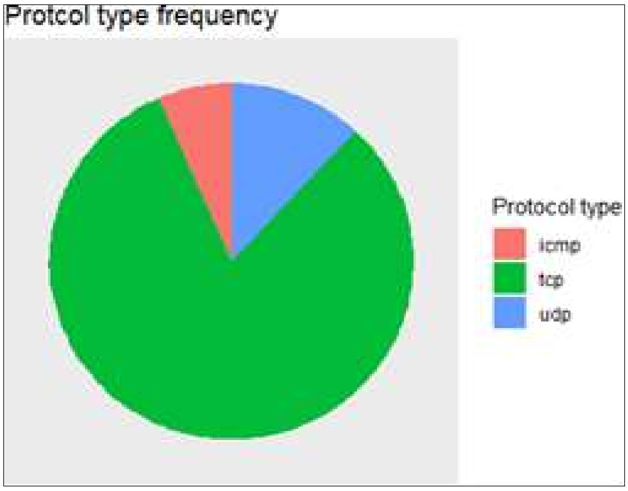


Fig 2: Protocol type frequency

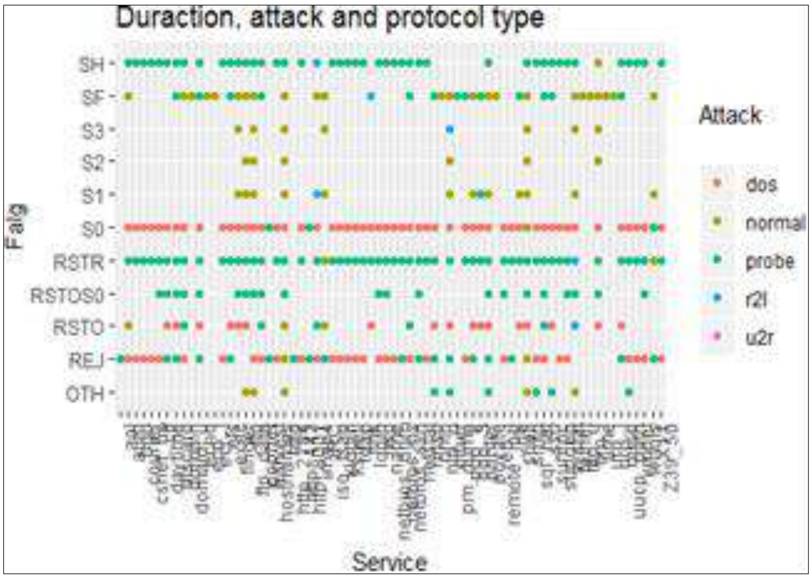
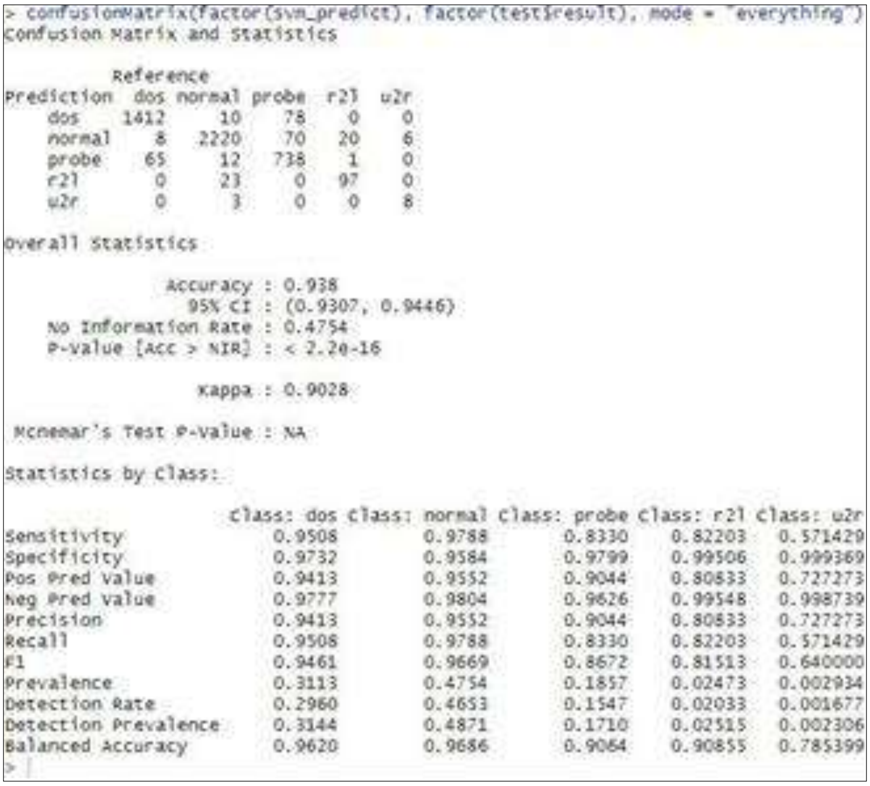


Fig 3: Correlation between duration, attack and protocol type

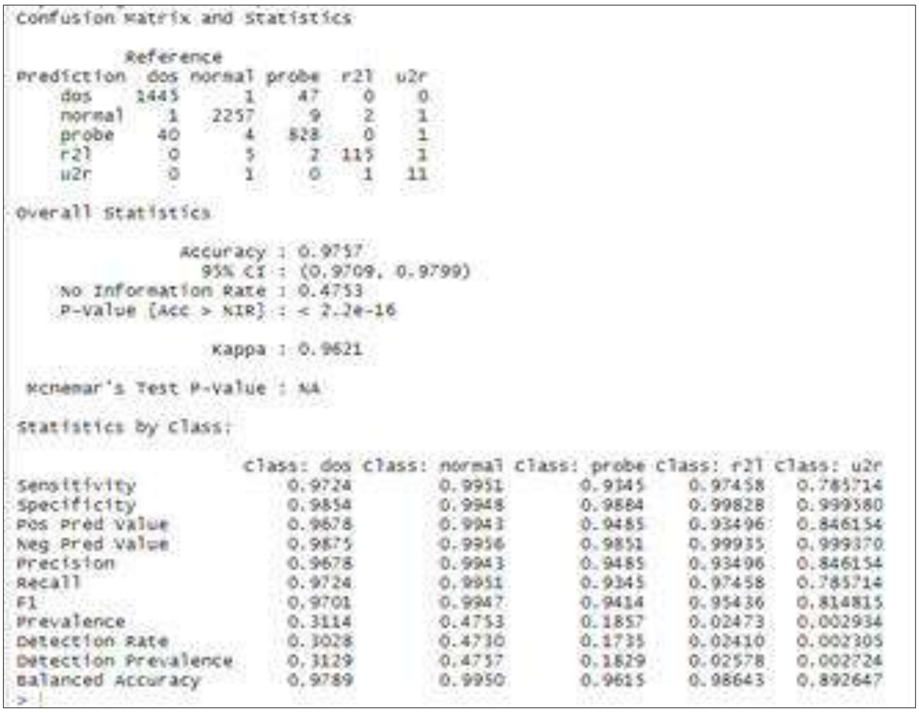
The next process is to use this dataset and put it across machine-learning algorithms that might give good results by correctly classifying the instances. First we can train the SVM model which is an single learning classifier means the output or prediction is taken from single classifier and the testing result of the SVM classifier on test dataset are shown in figure 4.



```
> confusionMatrix(factor(svm_predict), factor(test$result), mode = "everything")
Confusion Matrix and Statistics

          Reference
Prediction  dos normal probe  r2l  u2r
    dos    1412     10    78    0    0
    normal    8   2220    70   20    6
    probe    65     12   738    1    0
    r2l       0     23     0   97    0
    u2r       0      3     0    0    8

Overall Statistics

               Accuracy : 0.938
                 95% CI : (0.9307, 0.9446)
    No Information Rate : 0.4754
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9028

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: dos Class: normal Class: probe Class: r2l Class: u2r
Sensitivity              0.9508        0.9788       0.8330    0.82203   0.571429
Specificity              0.9732        0.9584       0.9799    0.99506   0.999369
Pos Pred Value           0.9413        0.9552       0.9044    0.80833   0.727273
Neg Pred Value           0.9777        0.9804       0.9626    0.99548   0.998739
Precision                0.9413        0.9552       0.9044    0.80833   0.727273
Recall                   0.9508        0.9788       0.8330    0.82203   0.571429
F1                       0.9461        0.9669       0.8672    0.81513   0.640000
Prevalence               0.3113        0.4754       0.1857    0.02473   0.002934
Detection Rate           0.2960        0.4653       0.1547    0.02033   0.001677
Detection Prevalence     0.3144        0.4871       0.1710    0.02515   0.002306
Balanced Accuracy        0.9620        0.9686       0.9064    0.90855   0.785399
> |
```

Fig 4: Confusion Matrix and Statistics of SVM model

In this SVM is gives 93.8% of overall accuracy but the accuracy is given by single classifier. In proposed we can introduce XGBoost learning in machine learning in which they combine the decisions from multiple models to improve the overall performance. Ensembling is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model. In the above example, the way we combine all the predictions together. We can build a model based on XGBoost learning and trained the model and the testing result are shown in figure5.



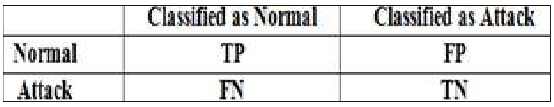
```
Confusion Matrix and Statistics

          Reference
Prediction  dos normal probe  r2l  u2r
    dos    1445      1    47    0    0
    normal    1   2257     9    2    1
    probe    40      4   828    0    1
    r2l       0      5     2  115    1
    u2r       0      1     0    1   11

Overall Statistics

               Accuracy : 0.9757
                 95% CI : (0.9709, 0.9799)
    No Information Rate : 0.4753
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9621

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: dos Class: normal Class: probe Class: r2l Class: u2r
Sensitivity              0.9724        0.9951       0.9345    0.97458   0.785714
Specificity              0.9854        0.9948       0.9884    0.99828   0.999580
Pos Pred Value           0.9678        0.9943       0.9485    0.93496   0.846154
Neg Pred Value           0.9875        0.9956       0.9851    0.99935   0.999370
Precision                0.9678        0.9943       0.9485    0.93496   0.846154
Recall                   0.9724        0.9951       0.9345    0.97458   0.785714
F1                       0.9701        0.9947       0.9414    0.95436   0.814815
Prevalence               0.3114        0.4753       0.1857    0.02473   0.002934
Detection Rate           0.3028        0.4730       0.1735    0.02410   0.002305
Detection Prevalence     0.3129        0.4757       0.1829    0.02578   0.002724
Balanced Accuracy        0.9789        0.9950       0.9615    0.98643   0.892647
> |
```

Fig 5: Confusion Matrix and Statistics of XGBoost model

5.1 Performance Measure

We used accuracy, which are derived using confusion matrix.

Table 3: Confusion Matrix



	Classified as Normal	Classified as Attack
Normal	TP	FP
Attack	FN	TN

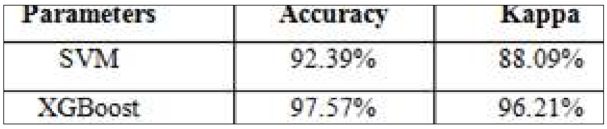
Where

TN -Instances correctly predicted as non-attacks. FN -Instances wrongly predicted as non-attacks. FP -Instances wrongly predicted as attacks.

TP -Instances correctly predicted as attacks.

$$\text{Accuracy} = \frac{\text{Number of samples correctly classified in test data}}{\text{Total number of samples in test data}}$$

Table 4: Performance Measure of models



Parameters	Accuracy	Kappa
SVM	92.39%	88.09%
XGBoost	97.57%	96.21%

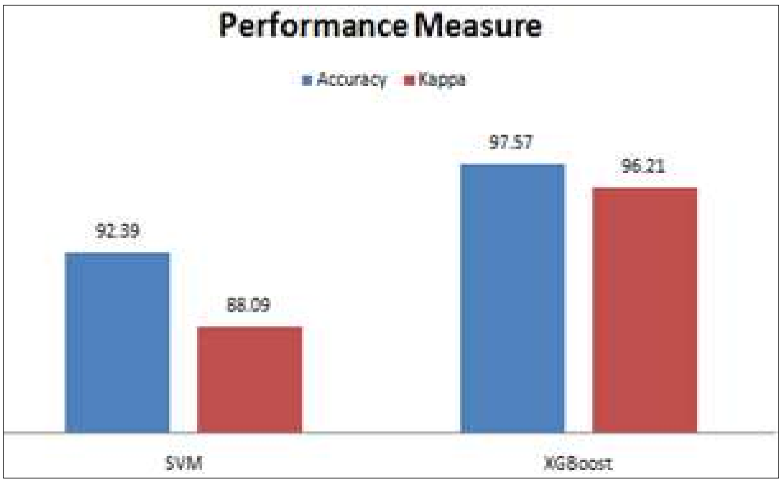
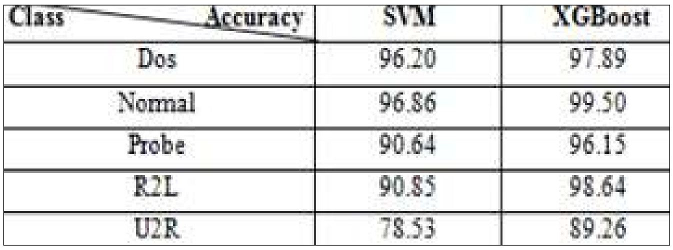


Fig 6: Performance Measure

In these we can also compute the performance on the basis of class wise, Table shows the accuracy based on different class.

Table 5: Class wise performance measure



Class Accuracy	SVM	XGBoost
Dos	96.20	97.89
Normal	96.86	99.50
Probe	90.64	96.15
R2L	90.85	98.64
U2R	78.53	89.26

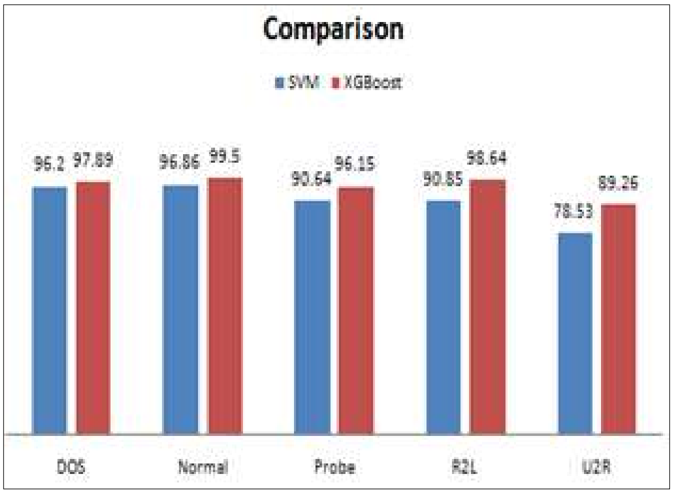


Fig 7: Class wise comparison

6. Conclusion

Traditional IDS suffer from different problems that limit their effectiveness and efficiency. In contrast machine learning techniques are promising approaches for intrusion detection. Classification of Intrusion Detection System (IDS) with NSL-KDD99 dataset applies classification technique of SVM and XGBoost Learning method.

After testing the results of different classification results we can say that XGBoost learning classifiers gives better performance in terms of accuracy as compared to SVM Classifier.

7. References

1. Goyal, Abhilash, Gupta, Divyansh, Intrusion Detection and Prevention in Software Defined Networking in IEEE 2018.
2. Rohit Kumar Singh Gautam, Er. Amit Doegar. An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms in IEEE 2018.
3. Kinam Park, Youngrok Song, Yun-Gyung Cheong. Classification of Attack Types for Intrusion Detection Systems using a Machine Learning Algorithm in IEEE 2018.
4. Bay SD, Kibler DF, Pazzani MJ, Smyth P. Theucikdd archive of large data sets for data mining research and experimentation, ‖SIGKDD Explorations 2000, 2(81).
5. Jain, Jay Kumar. Secure and energy-efficient route adjustment model for internet of things. Wireless Personal Communications 108.1 2019, 633-657.