I. J. Computer Network and Information Security, **2019, 7, 37-44**
Published Online July 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2019.07.06

Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier

N Manju
Department of Information Science and Engineering,
Sri Jayachamarajendra College of Engineering, Mysuru, 570006, India.
E-mail: manjun007@sjce.ac.in

B S Harish and V Prajwal
Department of Information Science and Engineering,
JSS Science and Technology University, Mysuru, 570006, India.
E-mail: bsharish@jssstuniv.in and prajwalv.94@gmail.com

Received: 04 May 2019; Accepted: 28 May 2019; Published: 08 July 2019

Abstract—Identification and classification of internet traffic is most important in network management to ensure Quality of Service (QoS). However, existing machine learning models tend to produce unsatisfactory results when applied with imbalanced datasets involving multiple classes. There are two reasons for this: the models have a bias towards classes which have more samples and they also tend to predict only the majority class data as features of the minority class are often treated as noise and therefore ignored. Thus, there is a high probability of misclassification of the minority class compared with the majority class. Therefore, in this paper, we are proposing an ensemble feature selection based on the tree approach and ensemble classification model using XGboost to enhance the performance of classification. The proposed model achieves better classification accuracy compared to other tree based classifiers.

***Index Terms*—**Identification, Classification, Feature Selection, Internet Traffic.

I. Introduction

The Internet traffic is made up of flows from various applications throughout the heterogeneous network. Most of these application parameters are unique in the network. Basic understanding of these applications and protocols is essential for any network management tool to manage traffic. The main activities of network management are fault diagnosis, anomaly detection, capacity provisioning and planning, application performance and providing quality of service (QoS). Due to easy availability of broadband internet connection and improvement in the quality of service, users are highly inclined to use the broad range of available services over the internet, few of which are Voice over Internet Protocol (VoIP), Internet banking, e-commerce, P2P systems and many more. This leads to complexity in the behavior of the internet beyond the understanding of the typical user [1]. Thus, Internet Service Providers (ISP) must pay more attention towards the complexity of the behavior of network.

Internet traffic classification has the ability to solve various network management issues for Internet Service Providers. Traffic classification is an important part of an automated system, where it detects intrusion [2,3] the patterns of Denial of Service (DoS), network resource reallocation based on customer demand [4], lawful interception [5] etc.

Traditional internet traffic classification depends on inspection of packets on TCP or UDP port numbers and payload-based approach. Each technique suffers from its own limitations. In port based, internet traffic can be classified based on port numbers. However, today's applications assign port numbers dynamically, hence port based classification is not reliable anymore. On the other hand, payload based classification involves finding patterns of application by analyzing packet content in the header. However, it fails to identify patterns of applications, whether the data is encrypted or violates the privacy of users [6]. To overcome these drawbacks, statistical approach is used to identify the application based on characteristics such as flow duration, flow idle time, packet length etc. These properties are unique for some classes of applications which differ from each other [7,8]. To solve these classification problems, applying machine learning algorithms has become one of the popular areas of research. Various machine learning based internet traffic classification methods are proposed and significant results have been achieved [9].

The rest of this paper is organized as follows: Literature survey is presented in Section 2 and methodology is presented in Section 3. In section 4, experimental settings are illustrated along with results and analysis. Paper is concluded in Section 5 and future scope is given.

II. Literature Survey

IP traffic classification is an essential part of traffic management such as identifying abnormal behaviour, application prioritization and delivery of QoS. However, port based and payload based classifications are inefficient approaches due to drawbacks as stated earlier. Hence, most of the internet traffic classification is done based on statistical flow features [10,11]. The performances of various machine learning algorithms are demonstrated using Naive Bayes, C4.5, Bayesian Network and Naive Bayes Tree [12]. The results show that C4.5 achieves faster classification speed and that Naive Bayes Kernel is slower compared with other algorithms. Further, the classification achieved has an average accuracy of 95% with all algorithms. In [13], supervised machine learning algorithms such as Bayesian Networks, Decision Tree and Multilayer perceptrons are used for evaluation and comparison. Overall results show that decision tree is most suitable for achieving traffic classification using multilayer perceptron but its accuracy is lower. Support Vector Machine (SVM) is proposed in [14] to experiment with biased and unbiased data samples on both training and testing. The best results are achieved for biased training and testing samples compared with unbiased samples. Performance of machine learning algorithms such as C5.0, Adaboost and Genetic programming is used to identify Skype VoIP encrypted data in [15]. The result shows that uniform sampling using random selection is appropriate for achieving better accuracy. Improved Security Information and Event Management (ISIEM) is used to identify Skype traffic using ad-hoc developed by enhanced probe. It includes classification engine which is influenced by machine learning to expose encrypted VoIP Skype traffic. Types of classification engines include J48, Logistic and Bayesian Networks [16]. Multi Objective Evolutionary Fuzzy Classifiers are proposed in [17]. Proposed methods are based on Fuzzy Rule Based Classifiers (FRBC) and result appears satisfactory. Identification of applications rather than categories is proposed using J48, Random Forest, K-NN and Bayes Net using complete 111 features of UNBISCX standard dataset [18]. The results show an accuracy of 93.44% using KNN for ISCX datasets while Random Forest achieved an accuracy of 90.87% for internal datasets. During the next set of experimentation, reducing the number of set of features from 111 to 12 features increased accuracy by 2% for the internal dataset. Extreme Learning Machine (ELM) approach is used to classify the internet traffic; Kernel Based Extreme Learning Machine is applied on Cambridge dataset and developed into software based on genetic algorithm to select the parameters. The result achieved an accuracy of 95% [19]. Multi-class imbalance is also one of the problems while using machine learning approaches for training and testing the samples. To resolve this problem, cost-sensitive method based on Flow rate based Cost Matrix (FCM) and Weighted Cost Matrix (WCM) is proposed in [20]. Results show that WCM performs well when compared with FCM in terms of stability. Imbalanced Data Gravitation-based Classification (IDGC) based model is developed to fix data imbalance problem using five standard and four imbalanced algorithms for experimentation. The result shows that conventional classification models are not as effective as IDGC to correct imbalanced internet traffic data. On the other hand, C4.5CS also performs equally well when compared with IDGC as presented in [21].

Feature selection method plays an important role in classification using machine learning approaches. This method is used to identify subset of relevant features and remove redundant ones from each of the feature subset. In machine learning, various methods are proposed to solve classification problems, some of which are discussed in this section. Extraction of real-time feature subset is proposed in [22] and performance is evaluated using various machine learning algorithms based on decision tree algorithms. Hybrid approach is proposed based on discretization, filtering and classification methods [23]. The result shows that, hybrid method achieves better performance. Weighted Symmetrical Uncertainty Area Under ROC Curve hybrid approach is presented in [24]. Further, Selecting Robust and Stable Feature (SRSF) is applied to evaluate the internet traffic data for various datasets. Experimentation results show that the proposed method gives better result using C4.5 in terms of accuracy and speed. Balanced Feature Selection (BFS) method is proposed in [25] and compared with other feature selection algorithms such as Information Gain (IG), Chi-Squared based, Fast Correlation Based Filter (FCBF), Correlation-based Feature Selection (CFS) and CON (Consistency-based selection). Result shows improvement in the g-mean of 15.29 compared to other methods mentioned above. In [26], bias coefficient for BFS is compared with FCBF. BFS gives better result than FCBF using Naive Bayes classifier. MAUC is the metric which is an improved version of AUC (Area under Curve). The success rate for MAUC is 93% and accuracy achieved is 90%. Feature selection based on rough set theory is proposed in [27] which reduce the feature set from 6 to 10 features. Using these features set, Bayesian network is used to improve the classification accuracy.

Also, identifying the metrics required to provide quality result is important in any application. In a similar way, authors in [28] proposed metrics such as goodness, stability and similarity to select features from the feature set. These metrics are applied on various feature selection methods on all 10 datasets of Cambridge University. The result predicts different values for different datasets. Final result shows that a combined feature selection technique gives better result than using individual feature selection methods. Hybrid approach called Global Optimization Approach (GOA), based on multi-criterion fusion for optimal and information theoretic for stable features are proposed in [29]. Proposed method is compared with various machine learning algorithms which improve classification accuracy. Class Oriented Feature Selection (COFS) is proposed in [30] and the proposed method is compared using different machine learning algorithms. C4.5 algorithm achieves better result compared with all other learning algorithms. Multifractal feature extraction and selection using Wavelet Leaders Multifractal Formalism (WLMF) is proposed in [31]. Support Vector

Machine is used for classification of both TCP and UDP flows. Overall accuracy achieved by TCP flow is 95.67% and by UDP flow is 97.67%. Robust feature selection method is proposed in [32] which select 3 to 4 features out of 17 features present in the dataset, using mutual information analysis. Learning of Decomposable Models with Limited Cycle Size (LDMLCS) is proposed in [33] to extract dependencies among features.

From the literature survey, we found several issues such as multiclass imbalance problem, scope for improvement in feature selection process and no common results achieved for different types of datasets. However, the identified issues provide avenues for further experimentation.

III. Methodology

Internet traffic identification and classification has its own importance in network management. Study and evaluation of the process involves various stages during the experimentation. Most important part of the experimentation is based on data collection which can be done from various sources of repository. The next stage of process starts from sampling and pre-processing the collected data. Further, we apply feature subset selection approach which plays a vital role in classification to evaluate the performance of the model.

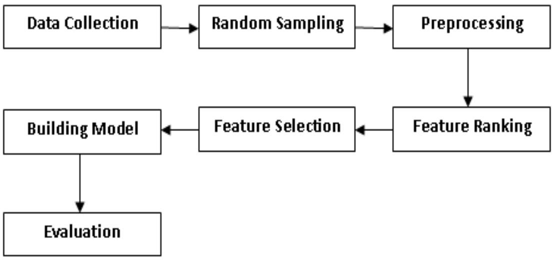

Fig.1. Block diagram of proposed model

Fig.1 shows the block diagram of the proposed model applied for our experimentation. The most important part of experimentation starts from collecting the dataset. Next, we use random sampling to get derived dataset01. Further, we apply pre-processing method followed by feature ranking based on tree structure. Then, we select the features based on the classification accuracy of the proposed XGBoost classification model as shown in fig 2. Finally, we evaluate the result which is discussed later.

A. Pre-processing

Use **Standardization**: The data present in the different columns correspond to different scales and have different units. They were scaled down to a standard range with mean value (μ) =0 and standard deviation (σ) =1.The linear transformed values are called z-scores which are computed using the equation shown below.

Z -score:

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

Where,

$'\mu'$ is the mean given by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{x}_i) \tag{2}$$

$'\sigma'$ is the standard deviation given by:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{x}_i - \mu)^2} \tag{3}$$

B. Feature Ranking

Feature selection method plays a key role in various classification tasks. It helps in enhancing the efficiency and accuracy of machine learning algorithms by selecting small subset from more number of features presented in the dataset. This process helps in identifying more discriminating features and removing redundant and irrelevant features. The feature which carries no information about the various classes is said to be irrelevant feature. On the other hand, if the feature has high correlation with other features and decreases the accuracy, it is said to be redundant feature.

In our work, we evaluate features based on feature ranking in which each feature is given a rank based on the weight associated with it. The weight of the feature corresponds to the number of times it appears in the tree. Feature defines the structure of the tree; therefore choosing the right features will result in better tree structure. Features are then placed in the decreasing order of their ranking.

The original features are arranged in the order of $x_1, x_2, x_3, x_4, ...x_n$. Upon applying the feature ranking, the features are ordered in the decreasing order of their weight $x_1', x_2', x_3', x_4', ...x_n'$. Here, x_1' corresponds to the feature of highest importance and x_n' corresponds to the feature of least importance.

C. Feature Selection

The result of feature ranking is ranked features, which build the model by adding each feature incrementally and recording the accuracy. The accuracy becomes almost persistent from feature 8 to 102. From feature 103, it drops down to 96% (approximately). We found that the change in accuracy was negligible compared to the number of features added. If more number of features is added, the model becomes more complex. Hence the number of features is restricted to 8. Fig 2 shows the accuracy for the first 117 features and fig 3 shows the features chosen based on the result obtained from fig 2.

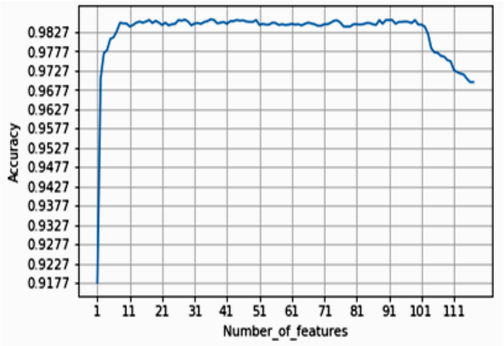

Fig.2. Accuracy of first 117 features

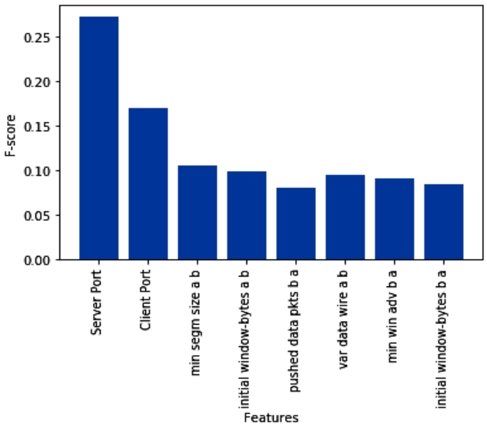

Fig.3. Features chosen based on obtained accuracy

D. *Model Building*

Extreme Gradient Boosting, known as XGBoost, is a scalable tree boosting system which incorporates efficiency and memory resources**.** Considering the nature of the dataset, it is important to improve the accuracy in the classes having fewer samples. Hence, incorporating the boosting technique fits the objective of the problem. Our proposed model is built using XGBoost algorithm where we construct multiple new models and combine them sequentially to form a final model until the error gets minimized and the accuracy becomes stable.

In machine learning algorithm, a major challenge is to build a highly reliable classifier model which has the ability to distinguish target applications based on effective feature set.

The data is divided into 60% training and 40% testing and is stratified to handle the class imbalance in both test and train. The model parameters are empirically set as follows:

1. Maximum depth of tree = 4, which reduces the length of the decision path as shown in fig 4
2. Learning rate = 0.1, which is learning rate of boosting
3. Evaluation_metric = log loss

Log Loss quantifies the accuracy of a classifier by penalising false classifications (i.e., decrease in log loss will increase the accuracy)

Choosing accurate number of estimators and depth of the tree is most important. When we build the model in tree based algorithms, it decides the complexity of the model. The depth of the tree and number of estimators are computed together against the logloss function. From this function, depth of the tree was computed to be 4 and number of estimators to be 300, after conducting many trails empirically. The numbers of estimators is computed using an ensemble and the difference of errors in the model is captured using a logloss function as shown in the equation (4).

$$\text{logloss function} = -\sum_{i=1}^{C} b_{ji} \log(\mathrm{P}_{ji}) \tag{4}$$

C = Number of classes.

b = Binary indicator (Checks if the class label i is the correct classification for observation j. If yes, then the value will be ‘1’ else ‘0’)

Where, i =iterative variable upto number of classes C, j =predicted label, P = predicted probability observation of j belonging to class i.

The prediction scores of each individual estimator are added to get a final score using the following equation (5).

$$y_k = \sum_{i=1}^{E} f_i(\mathrm{x}_\mathrm{k}) \tag{5}$$

E = Number of estimators

f = Function which estimates set of all possible classifications

y = Prediction score for k^{th} sample

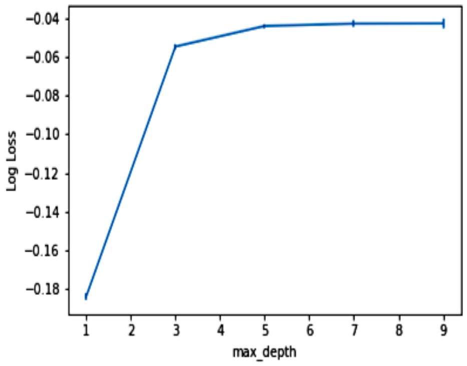

Fig.4. Depth of the tree using logloss function

Identifying the depth of the tree and number of estimators is estimated by computing the logloss function, the goal being to minimize estimator’s value. Fig 4 shows the logloss values versus depth of the tree computed independently. Likewise, fig 5 shows the logloss values versus number of estimators computed independently finding an approximation to these parameters can be achieved by considering both depth of the tree and number of estimators required. When they are considered together, it produces multiple combinations. The optimal depth of the tree is selected as 4 and number of estimators as 300 since it minimizes the loss as shown in the fig 6. The result of the various combinations is also shown in fig 6.

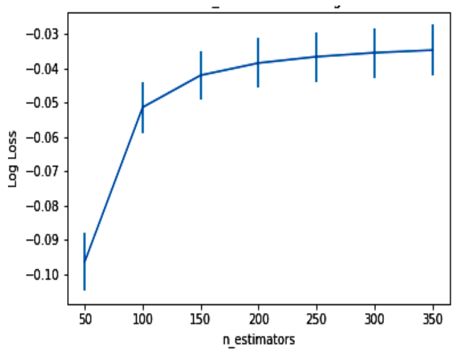

Fig.5. Estimators of the tree using logloss function

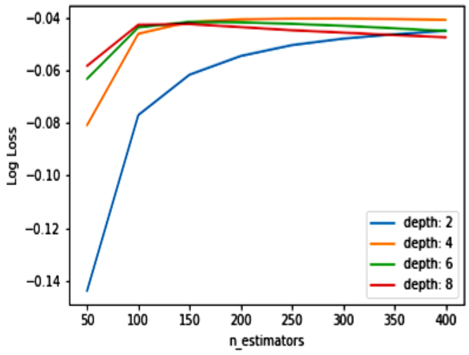

Fig.6. Values for depth of the tree and number of estimators using logloss function

E. *Evaluating the Model*

We have evaluated the model using 10-fold stratified cross validation. It was carried out to test the effectiveness of the model.

IV. Results and Discussion

This section is intended to evaluate the tree based ensemble feature selection approach and proposed XGboost model. In addition, we have compared with other existing machine learning algorithms which are of same family such as decision tree, random forest and adaboost. Proposed model is analyzed with and without feature selection approach.

A. *Dataset*

We have considered Cambridge datasets for our experimentation of which there are two variants. First dataset is a derived subset of Cambridge dataset called dataset01 which is normalized (less imbalance). This is compared with the Cambridge standard dataset called dataset02 which is highly imbalanced. It has 10 subsets of data captured over different periods of time in a day. We have combined all the flows and randomly sampled the data. The sample threshold for each class was set at maximum 3000 flows due to unavailable flows to majority of other classes. The flows from the class ‘Games’ and ‘Interactive’ were deleted due to fewer flows. The flow count for the rest of the classes can be seen in the table 1 below and it is considered as derived subset. The reason for deriving the dataset is to minimize the imbalance compared to high imbalanced datasets. On the other hand, we use standard dataset which is termed entry01, from Cambridge dataset for the experimentation. This dataset consists of highly imbalanced flows for various classes. We have referred to these two datasets as dataset01 and dataset02 respectively for our convenience.

Table 1. Datasets used for our experimentation

Class Name	Derived Subset **(Dataset01)**	Flow distribution Ratio in %	Standard Dataset **(Dataset02)**	Flow distribution Ratio in %
	Flow Count		**Flow Count**	
WWW	3000	12.55	18211	73.25
MAIL	3000	12.55	4146	16.67
FTP-DATA	3000	12.55	1319	05.30
FTP-CONTROL	3000	12.55	149	00.59
FTP-PASV	2688	11.24	43	00.17
DATABASE	2648	11.08	238	00.95
SERVICES	2099	08.78	206	00.82
P2P	2094	08.76	339	01.36
ATTACK	1793	07.50	122	00.49
MULTIMEDIA	576	02.41	87	00.34
Total Number Flows	**23898**	**~ 100.00**	**24860**	**~ 100.00**

B. *Experimentation of Dataset01*

In this section, we study the result obtained for proposed XGboost model using dataset01 and compare with other tree based algorithms such as decision tree, random forest and adaboost. Proposed model is evaluated using 248 and 8 features.

The result obtained for dataset01 using 248 features and 8 features is shown in table 2 and table 3 respectively. We computed Precision, Recall and F1-score and obtained value ‘1’ for the FTP-DATA class. Various other classes fall below ‘1’ as seen in table 2. On the other hand, DATABASE, FTP-CONTROL, FTP-DATA, FTP-PASV and MAIL classes evaluation metrics are all ‘1’.Other classes also show improvement in performance by being close to '1', using 8 features as shown in table 3. It is clear that the proposed model with 8 features gives better result in terms of Precision, Recall and F1-score of the performance metrics for the derived subset dataset01 compared to all 248 features.

Table 2. Evaluation metrics used for dataset01 using 248 features

Without feature selection of dataset01			
Categories of Classes	**Precision**	**Recall**	**F1-score**
ATTACK	0.93	0.91	0.92
DATABASE	0.97	0.93	0.95
FTP-CONTROL	0.97	0.97	0.97
FTP-DATA	1.00	1.00	1.00
FTP-PASV	0.96	0.97	0.96
MAIL	0.99	1.00	0.99
MULTIMEDIA	0.99	0.98	0.98
P2P	0.97	0.97	0.97
SERVICES	1.00	0.99	0.99
WWW	0.93	0.97	0.95

Table 3. Evaluation metrics used for dataset01 using 8 selected features

With feature selection of dataset01			
Categories of Classes	**Precision**	**Recall**	**F1-score**
ATTACK	0.97	0.86	0.91
DATABASE	1.00	1.00	1.00
FTP-CONTROL	1.00	1.00	1.00
FTP-DATA	1.00	1.00	1.00
FTP-PASV	1.00	1.00	1.00
MAIL	1.00	1.00	1.00
MULTIMEDIA	0.99	0.99	0.99
P2P	0.97	0.99	0.98
SERVICES	1.00	0.99	0.99
WWW	0.93	0.99	0.96

Table 4. Classification accuracy of various decision tree based algorithms on dataset01

Sl No	**Classifier Name**	**248 Features**	**8 Features**
		Accuracy in %	**Accuracy in %**
1	Decision Tree	83.26	87.63
2	Random Forest	86.78	92.27
3	Adaboost	91.53	94.59
4	**XGBoost**	**96.97**	**98.51**

The overall accuracy of the proposed model is compared with tree based algorithms such as decision tree, random forest and adaboost as shown in table 4. From the obtained result, it is clear that, proposed XGboost model outperforms other tree based classifiers for dataset01 with 98.51% accuracy.

C. Experimentation of Dataset02

In this section, we study the result obtained for proposed XGboost model using dataset02 and compare with other tree based algorithms such as decision tree, random forest and adaboost. Proposed model is evaluated using 248 and 8 features.

Table 5. Evaluation metrics used for dataset02 using 248 features

Without feature selection of dataset02			
Categories of Classes	**Precision**	**Recall**	**F1-score**
ATTACK	0.43	0.56	0.49
DATABASE	0.87	1.00	0.93
FTP-CONTROL	0.88	0.95	0.93
FTP-DATA	1.00	1.00	1.00
FTP-PASV	0.73	0.78	0.75
MAIL	1.00	0.99	0.99
MULTIMEDIA	0.88	0.36	0.51
P2P	0.97	0.87	0.92
SERVICES	1.00	0.99	0.99
WWW	1.00	1.00	1.00

Table 6. Evaluation metrics used for dataset02 using 8 selected features

With feature selection of dataset02			
Categories of Classes	**Precision**	**Recall**	**F1-score**
ATTACK	0.56	0.56	0.56
DATABASE	1.00	1.00	1.00
FTP-CONTROL	1.00	1.00	1.00
FTP-DATA	1.00	1.00	1.00
FTP-PASV	0.90	0.66	0.76
MAIL	1.00	1.00	1.00
MULTIMEDIA	0.93	1.00	0.96
P2P	0.97	0.99	0.98
SERVICES	1.00	0.99	0.99
WWW	1.00	1.00	1.00

The results for dataset02 using all 248 features and 8 features are shown in tables 5 and table 6 respectively. From the result, we can observe that there is improvement in performance of most of the classes such as DATABASE, FTP-CONTROL and MAIL. However, there is no change in the performance of SERVICES class.

Table 7. Classification accuracy of various decision tree based algorithms for dataset02

Sl No	**Classifier Name**	**248 Features**	**8 Features**
		Accuracy in %	**Accuracy in %**
1	Decision Tree	79.39	82.26
2	Random Forest	81.17	86.68
3	Adaboost	83.37	87.22
4	**XGBoost**	**87.48**	**93.54**

The overall performance of the proposed model when compared with tree based algorithms for dataset02 is shown in table 7. From the obtained result, it is clear that the proposed XGboost model outperforms other tree based classifiers for dataset02 with an accuracy of 93.54%.

V. Conclusion

Internet traffic classification plays an important task in network monitoring and management. Classification using machine learning algorithms gives promising results. Since internet traffic classification is a multiclass problem, many of the existing machines learning models do not perform well. There are many tree based machine learning classification models and variety of feature selection methods that can be used for classification. The proposed model in this paper performs with accuracy of 98.51% and 93.54% in classifying dataset01 and dataset02 respectively using only 8 selected features. Reduction in the features also decreases the computational overhead. In future, we intend to handle the issue of imbalanced dataset to enhance the classification accuracy.

References

[1] Cho, K., Fukuda, K., Esaki, H., and Kato, A. “The impact and implications of the growth in residential user-to-user traffic”, in: *ACM SIGCOMM Computer Communication Review*, Vol. 36, No.4, pp. 207-218, 2006.

[2] Roesch, M., “Snort: Lightweight intrusion detection for networks”, in: *Lisa,* Vol. 99, No. 1, pp. 229-238, 1999.

[3] Paxson, V., “Bro: a system for detecting network intruders in real-time”, *Computer networks*, 31(23-24), pp. 2435-2463, 1999.

[4] Stewart, L., Armitage, G., Branch, P., and Zander, S., “An architecture for automated network control of QoS over consumer broadband links”, 2005.

[5] Baker, F., Foster, B., and Sharp, C., “Cisco architecture for lawful intercept in IP networks”, (No. RFC 3924), 2004.

[6] Kim, H., Claffy, K. C., Fomenkov, M., Barman, D., Faloutsos, M., and Lee, K, “Internet traffic classification demystified: myths, caveats, and the best practices”, in: *Proceedings of the 2008 ACM CoNEXT conference*, pp. 11, 2008.

[7] Paxson, V., “Empirically derived analytic models of wide-area TCP connections”, *IEEE/ACM Transactions on Networking (TON)*, Vol. 2, No. 4, pp. 316-336, 1994.

[8] Dewes, C., Wichmann, A., and Feldmann, A., “An analysis of Internet chat systems”, in: *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pp. 51-64, 2003.

[9] WANG, R. Y., Zhen, L. I. U., and ZHANG, L., “Method of data cleaning for network traffic classification”, *The Journal of China Universities of Posts and Telecommunications*, Vol. 21, No. 3, pp.35-45, 2014.

[10] Lin, P., Yu, X. Y., Liu, F., and LEI, Z. M., “A network traffic classification algorithm based on flow statistical characteristics”, *Journal of Beijing University of Posts and Telecommunications*, Vol. 31, No. 2, pp. 15-19, 2008.

[11] Min, L. I. U. Q. L. I. U. Z., “Study on Internet Traffic Classification Using Machine Learning”, *Computer Science*, 12, 008.2010.

[12] Williams, N., Zander, S., & Armitage, G., “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification”, in: *ACM SIGCOMM Computer Communication Review*, Vol. 36, No. 5, pp. 5-16, 2006.

[13] Soysal, M., and Schmidt, E. G., “Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison”, *Performance Evaluation*, Vol. 67, No. 6, pp. 451-467, 2010.

[14] Yuan, R., Li, Z., Guan, X., and Xu, L.. “An SVM-based machine learning method for accurate internet traffic classification”, *Information Systems Frontiers*, Vol. 12, No. 2, pp.149-156, 2010.

[15] Alshammari, R., and Zincir-Heywood, A. N., “Identification of VoIP encrypted traffic using a machine learning approach”, *Journal of King Saud University-Computer and Information Sciences*, Vol. 27, No. 1, pp.77-92, 2015.

[16] Di Mauro, M., and Di Sarno, C., “Improving SIEM capabilities through an enhanced probe for encrypted Skype traffic detection”, *Journal of Information Security and Applications*, Vol. 38, pp. 85-95, 2018.

[17] Ducange, P., Mannarà, G., Marcelloni, F., Pecori, R., amd Vecchio, M., “A novel approach for internet traffic classification based on multi-objective evolutionary fuzzy classifiers”, in: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1-6, 2017.

[18] Yamansavascilar, B., Guvensan, M. A., Yavuz, A. G., and Karsligil, M. E., “Application identification via network traffic classification”, in: *International Conference on Computing, Networking and Communications (ICNC)*, pp. 843-848, 2017.

[19] Ertam, F., and Avcı, E., “A new approach for internet traffic classification: GA-WK-ELM”, *Measurement*, Vol. 95, pp. 135-142, 2017.

[20] Zhen, L. I. U., and Qiong, L. I. U., “Studying cost-sensitive learning for multi-class imbalance in Internet traffic classification”, *The Journal of China Universities of Posts and Telecommunications*, Vol. 19, No. 6, pp. 63-72, 2012.

[21] Peng, L., Zhang, H., Chen, Y., and Yang, B., “Imbalanced traffic identification using an imbalanced data gravitation-based classification model”, Computer *Communications*, Vol. 102, pp. 177-189, 2017.

[22] Zhao, J. J., Huang, X. H., Qiong, S. U. N., and Yan, M. A., “Real-time feature selection in traffic classification”, *The Journal of China Universities of Posts and Telecommunications*, Vol. 15, pp. 68-72, 2008.

[23] Bolon-Canedo, V., Sanchez-Marono, N., and Alonso-Betanzos, A., “Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset”, *Expert Systems with Applications*, Vol. 38, No. 5, pp. 5947-5957, 2011.

[24] Zhang, H., Lu, G., Qassrawi, M. T., Zhang, Y., and Yu, X., “Feature selection for optimizing traffic classification”, *Computer Communication*, Vol. 35, No. 12, pp. 1457-1471, 2012.

[25] Liu, Z., and Liu, Q., “Balanced feature selection method for Internet traffic classification”, *IET networks*, Vol. 1, No. 2, pp. 74-83, 2012.

[26] Zhen, L., and Qiong, L., “A new feature selection method for internet traffic classification using ml”, *Physics Procedia*, Vol. 33, pp. 1338-1345, 2012.

[27] Sun, M., Chen, J., Zhang, Y., and Shi, S, A new method of feature selection for flow classification, Physics Procedia, 24, 2012, pp. 1729-1736.

[28] Fahad, A., Tari, Z., Khalil, I., Habib, I., and Alnuweiri, H., “Toward an efficient and scalable feature selection approach for internet traffic classification”, *Computer Networks*, Vol. 57, No. 9, pp. 2040-2057, 2013.

[29] Fahad, A., Tari, Z., Khalil, I., Almalawi, A., and Zomaya, A. Y., “An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion”, *Future Generation Computer Systems*, Vol. 36, pp. 156-

169, 2014.

[30] Liu, Z., Wang, R., Tao, M., and Cai, X., “A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion”, *Neurocomputing*, Vol. 168, pp. 365-381, 2015.

[31] Shi, H., Li, H., Zhang, D., Cheng, C., and Wu, W., “Efficient and robust feature extraction and selection for traffic classification”, *Computer Networks*, Vol. 119, pp. 1-16, 2017.

[32] Shafiq, M., Yu, X., and Wang, D., “Robust Feature Selection for IM Applications at Early Stage Traffic Classification Using Machine Learning Algorithms”, in: *IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 239-245, 2017.

[33] Ghofrani, F., Keshavarz-Haddad, A., and Jamshidi, A., “A new probabilistic classifier based on decomposable models with application to internet traffic”, *Pattern Recognition*, Vol. 77, pp. 1-11, 2018.

Authors’ Profiles

N. Manju: He obtained his B.E in Computer Science and Engineering in the year 2005 and M.Tech in Computer Network Engineering in 2008 from Visvesvaraya Technological University, Belagavi, Karnataka, India. He is presently working as an Assistant Professor in the Department of Information Science & Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka, India. He is a Life Member of ISTE. His area of interest includes Machine Learning and Computational Intelligence.

B. S. Harish: He obtained his B.E in Electronics and Communication (2002), M.Tech in Networking and Internet Engineering (2004) from Visvesvaraya Technological University, Belagavi, Karnataka, India. He completed his Ph.D. in Computer Science (2011); thesis entitled “Classification of Large Text Data” from University of Mysore. He is presently working as an Associate Professor in the Department of Information Science & Engineering, JSS Science & Technology University, Mysuru. He was invited as a Visiting Researcher to DIBRIS - Department of Informatics, Bio Engineering, Robotics and System Engineering, University of Genova, Italy from May- July 2016. He delivered various technical talks in National and International Conferences. He has invited as a resource person to deliver various technical talks on Data Mining, Image Processing, Pattern Recognition, Soft Computing. He is also serving and served as a reviewer for National, International Conferences and Journals. He has published more than 50 International reputed peer reviewed journals and conferences proceedings. He successfully executed AICTE-RPS project which was sanctioned by AICTE, Government of India. He also served as a secretary, CSI Mysore chapter. He is a Member of IEEE (93068688), Life Member of CSI (09872), Life Member of Institute of Engineers and Life Member of ISTE. His area of interest includes Machine Learning, Text Mining and Computational Intelligence.

V. Prajwal: He obtained his B.E in Information Science and Engineering in 2016 from Electronics and Communication (2002), Visvesvaraya Technological University, Belagavi and currently perusing M.Tech in Data Science from JSS Science and Technology University, Mysuru, Karnataka, India. His area of interest includes Machine Learning, Computational Intelligence and Natural Language Processing.

How to cite this paper: N Manju, B S Harish, V Prajwal, "Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier", International Journal of Computer Network and Information Security(IJCNIS), Vol.11, No.7, pp.37-44, 2019. DOI: 10.5815/ijcnis.2019.07.06