# Performance Evaluation of Five Machine Learning Algorithms and Three Feature Selection Algorithms for IP Traffic Classification

Kuldeep Singh
University Institute of Engineering & Technology
Panjab University
Chandigarh, India

S. Agrawal
University Institute of Engineering & Technology
Panjab University
Chandigarh, India

## ABSTRACT
As volume of internet traffic over last couple of years due to drastic rise in number of internet users, the area of IP traffic classification has gained significant importance for various internet service providers and other public and private sector organizations. In today's scenario, traditional IP traffic classification techniques such as port number based and payload based techniques are rarely used because of their limitations of use of dynamic port number instead of well-known port number in packet headers and various cryptographic techniques which inhibit inspection of packet payload. In order to overcome these limitations, machine learning (ML) techniques are used for IP traffic classification. In this research paper, real time internet traffic dataset has been developed using packet capturing tool and then using three different feature selection algorithms: Correlation based, Consistency based and Principal Components Analysis based feature selection algorithms, reduced feature datasets have been developed. After that, five popular ML algorithms MLP, RBF, C4.5, Bayes Net and Naïve Bayes are used for IP traffic classification with these datasets. This experimental evaluation shows that C4.5 Decision Tree Algorithm is an efficient ML technique for IP traffic classification with reduction in number of features characterizing each internet application using Correlation based Feature Selection Algorithm.

## General Terms
IP Traffic Classification, Machine Learning, Internet Traffic.

## Keywords
MLP, RBF, C4.5, Bayes Net, Naïve Bayes.

## 1. INTRODUCTION
With rapid spread of internet all over the world, number of internet users increase at drastic rate and these users use various internet applications in their day to day life which leads to drastic increase in IP traffic. Various internet applications contributing in IP traffic are www, e-mail, Web Media, FTP data, P2P, instant messaging, VoIP etc. Therefore, IP Traffic Classification has gained sufficient importance for various internet service providers (ISPs) and other public and private sector organisations for various activities such as available bandwidth planning, fault diagnosis in network, analysis of quality of Service of any internet application or service, billing information about users who use a particular internet application, Lawful Interception of internet traffic data by certain government agencies in order to solve certain security related issues [1], [2].

Peer to peer (P2P) applications such as Bit Torrent, Emule, Kaaza etc are the major cause of rise in IP traffic which contribute to 80% of total volume of internet traffic [2]. Now a day, various multimedia websites such as Youtube which are actually audio and video streaming websites, contribute a major proportion in internet traffic. Common internet applications such www, e-mails and file transfer websites etc also have a significant share in this IP traffic. Most of the users in peak traffic hours use various messenger based applications such as Yahoo Messenger, Google Talk for audio and video calls and for instant messaging i.e. chatting which is again a major reason to rise in IP traffic.

Tradition IP traffic classification techniques are based upon direct inspection of packet headers and payloads such as port number based and payload based techniques [1], [3]. But in present time, use of these techniques is very rare. Port number based techniques become ineffective because of use of dynamic port number instead of well-known port number packet headers. While payload based technique becomes insignificant due to privacy policies of government and certain cryptographic techniques which are used to encrypt packet payload, inhibit inspection of IP traffic packets.

Current trends are use of Machine Learning (ML) techniques [1], [2], [11] for IP traffic classification which are based on training an algorithm using a dataset consisting of flow based statistical features, [7] and then using that trained algorithm for predicting classes of unlabelled test samples. These ML techniques overcome the limitations of inspection of packet headers and payloads. In present research paper, real time internet traffic dataset has been developed using packet capturing tool and then using three different feature selection algorithms: Correlation based, Consistency based Principal Components Analysis (PCA) based Feature Selection Algorithms, reduced feature datasets have also been developed. Then using these full feature and reduced feature datasets, five ML algorithms have been employed for IP traffic classification: Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF), C 4.5 Decision Tree Algorithm, Bayes Net Algorithm and Naïve Bayes Algorithm [11]. Performance of all these classifiers is analysed on the basis of classification accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications and number of features characterizing each internet application [1], [6].

The remaining paper is organised as follows: section 2 gives some information about related work done by various researchers in the field of IP traffic classification. Section 3 includes introductory information about all classifiers and

feature selection algorithms mentioned above. Section 4 gives overview of IP traffic dataset. Implementation and result analysis is given in section 5. Section 6 includes conclusions and future scope.

## 2. RELATED WORK

IP traffic classification is an emerging field in which is choice of number of researchers over last couple of years. For this research work, numbers of research papers have reviewed. Some of the previous work done in this field by some researchers is discussed as follows:

In [1], Nguyen et al. have given a brief introduction to various machine learning techniques for IP traffic classification. They have discussed port number based and payload based IP classification techniques and their limitations. Various machine learning techniques based upon clustering, supervised learning and hybrid approaches are explained briefly. Under clustering approach flow clustering using expectation maximization, automated application identification using AutoClass, TCP-based application identification using K- means, Identification of web and peer to peer in the network core techniques are explained. Under supervised learning approach, Statistical signature based technique using NN, LDA and QDA algorithms, Bayesian analysis techniques for, Real-time traffic classification using Multiple sub flows features, Generic algorithm based classification techniques etc are explained.

In [3], Runyuan Sun et al. have designed host based traffic collection platform to collect internet traffic of web, P2P and other applications. They have used three techniques for IP traffic classification such as Probabilistic neural network (PNN), RBF neural network and Support Vector Machine (SVM). They have concluded in their paper that PNN gives better performance as compared to other two networks. But this research work is limited to web and P2P applications only because they have not taken into account various other internet applications. Furthermore, by using other ML techniques there is chance of more improvement in classification performance in terms of accuracy and training time.

In [4], Singh and Agrawal have performed IP classification using RBF neural network and Back Propagation neural network. Performance of these networks is analyzed on the basis of classification accuracy, recall of individual applications, training time of networks and number of hidden layer neurons of the networks. In this research work, it is concluded that RBF neural network gives very good performance as compared to back propagation neural network at the cost of very high training time and computational complexity because at 1000 hidden layer neurons, RBF network gives 90.10 % classification accuracy. But training time is 432 minutes. Therefore, this approach is not suitable for online IP traffic classification because of its slow nature. There is still scope of further improvement in performance by using other ML algorithms for IP traffic classification.

In [5], Moore and Zeuv have described supervised Naïve Bayes classifier for IP traffic classification. They have utilized a dataset consisting of large number of data samples characterized by 248 flow features [7]. They find that their classifier's performance in terms of training time and classification accuracy reduces to large extent when redundant or irrelevant flow features are used. In order to overcome this problem, feature reduction is used to reduce number of features to the 12 most frequently used features. In the evaluation, Naive Bayes algorithm allows classification accuracy of flows to increase from 65% to over 95% due to feature reduction. But main problem is about dataset. The dataset, which is used in their research work, contains features related to port number and some other features related to packet header and payload. Thus, again, the problems related to packet payload and header inspection arise.

## 3. VARIOUS MACHINE LEARNING AND FEATURE SELECTION ALGORITHMS

In this research paper, five popular machine learning algorithms and three feature selection algorithms are used which are reported in different research papers to be performing well in most of the applications. These ML and feature selection algorithms are explained in brief as follows:

### 3.1 Multilayer Perceptron

Multilayer Perceptron (MLP), [8], [9], [18] also known as Back Propagation Neural Network, is a feed forward multilayer artificial neural network which is based upon extended gradient-descent based Delta learning rule, commonly known as Back Propagation rule. In this algorithm, error signal between desired output and actual output is being propagated in backward direction from output to hidden layer and then to input layer in order to train the algorithm and to use it as classifier.

The basic structure of MLP is shown in fig. 1. It consists of input layer having i neurons, hidden layer having j neurons and output layer having k neurons.

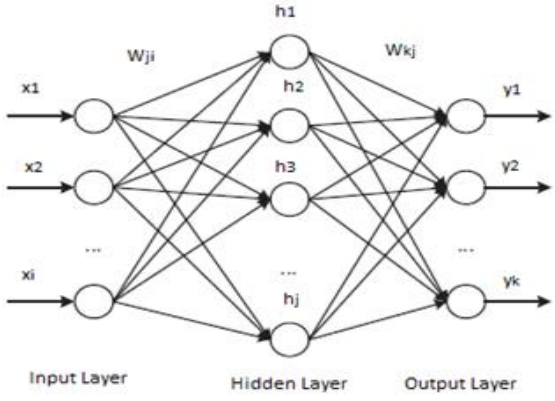


Fig 1: Multilayer Perceptron

In this research work, single hidden layer MLP is being used for IP traffic classification with learning rate of 0.3 and momentum term of 0.2 [12].

### 3.2 Radial Basis Function Neural Network

Radial Basis Function (RBF) Neural Network [4], [8], [19] is a multilayer feed forward artificial neural network in which radial basis functions are used as activation functions at each neuron in hidden layer. The output of this RBF neural network is weighted linear superposition of all these basis functions.

The basic structure of RBF neural network is shown in fig 2. In this network, weights are fixed for input-hidden layer

interconnections. While the weights for hidden-output layer interconnections are trainable. Each hidden layer neuron have a basis function $f_m(.)$.For any input vector X, the output of this network is given by following input-output mapping function as:

$$Y(X) = \sum_{i=0}^{m} Wi\, f(||X - Xi||) \quad (1)$$

Where $f(||X - Xi||)$ are M basis functions consisting of Euclidean distance between applied input X and training data point Xi.

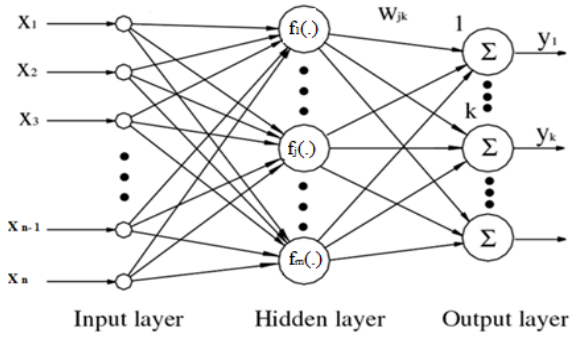


Fig 2: Radial Basis Function Neural Network

The commonly used basis function in RBF Algorithm is Gaussian function which is given as follows:

$$f(X) = \exp\left(-\frac{||X-\mu||}{2\sigma^2}\right) \quad (2)$$

Where μ is the Center and σ is spread constant which have direct effect on the smoothness of input-output mapping function Y(X).

In this research work, single hidden layer RBF algorithm has been used for IP traffic classification with number of center points in hidden layer equal to 5 and 2 for full feature dataset I and reduced feature datasets $II_{correlation}$ , $II_{consistency}$ and $II_{PCA}$ respectively [12].

## 3.3 C4.5 Algorithm

C4.5 is an ML algorithm which generates Univariate decision tree [10]. It is the extension of Iterative Dichotomiser 3 (ID3) algorithm which is used to find simple decision trees. C4.5 is also known as Statistical Classifier because its decision trees can be used for classification purpose.

C4.5 builds decision trees from a set of training data using the concept of information entropy. The training dataset consists of various training samples which are characterized by large number of features and also consists of target classes.

C4.5 chooses one feature of the data sample at each node of the tree which is used to split its set of samples into subsets having one class only for most of samples. This is based upon the criterion of normalized information gain which is obtained from choosing a feature for splitting the data. The feature with the highest normalized information gain is chosen to make the decision. After that, the C4.5 algorithm performs further subset splitting operation in same manner to make smallest subsets consisting of all the samples belonging to of single class only.

In this research work, C4.5 algorithm has been used for IP traffic classification with confidence factor of 0.25, minimum no. of instances per leaf equal to 2, no. of folds for pruning equal to 3 and seed used for randomizing the data, when error reduced pruning is used, equal to 1[11].

## 3.4 Bayes Net Algorithm

Bayes Net (Bayesian Network), [11], [13] also known as Belief Network, is a probabilistic graphical model which consists of acyclic graphical structure and conditional probability tables. This graphical structure is used to represent knowledge about some uncertain domain. In this model, each node represents a random variable and the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods.

Learning process in Bayesian Network take place in two steps: first learn a network structure, then learn the probability tables.

There are various approaches used for structure learning and in Weka tool, the following approaches are mainly considered:

- Local score metrics
- Conditional independence test
- Global score metrics
- Fixed structure

For each of these approaches, different search algorithms are implemented in Weka, such as hill climbing, simulated annealing and tabu search. Once a good network structure is identified, the conditional probability tables for each of the variables can be estimated.

In this research work, Bayes Net algorithm with simple estimator and K2 search algorithm has been used for IP traffic classification [10], [11].

## 3.5 Naïve Bayes Algorithm

A Naïve-Bayes (NB) ML algorithm [11], [14] is a simple structure consisting of a class node as the parent node of all other nodes. The basic structure of Naïve Bayes Classifier is shown in figure 3 in which C represents main class and a, b, c and d represents other feature or attribute nodes of a particular sample. No other connections are allowed in a Naïve-Bayes structure. Naïve-Bayes has been used as an effective classifier. It is easy to construct Naïve Bayes classifier as compared to other classifiers because the structure is given a priori and hence no structure learning procedure is required. Naïve-Bayes assumes that all the features are independent of each other.

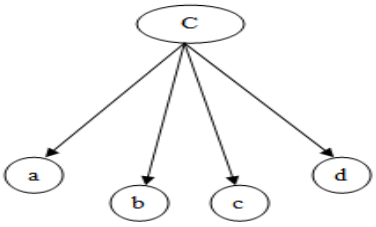


Fig 3: Naïve Bayes Classifier

Naïve-Bayes works very well over a large number of datasets, especially where the features used to characterize each sample are not properly correlated.

## 3.6 Correlation based Feature Selection Algorithm

Correlation based Feature Selection Algorithm, [15] is a popular algorithm which is used to identify and remove such irrelevant and redundant features as possible. Correlation based Feature Selection Algorithm is used to evaluate the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation between each other are preferred.

Correlation based Feature Selection Algorithm uses an evaluation procedure that examines the usefulness of individual features along with the level of inter-correlation among the features. High scores are assigned to subsets containing attributes that are highly correlated with the class and have low inter-correlation with each other. Concept of conditional entropy is taken in account in order to provide a measure of the correlation between features and class and between features. If H(X) is the entropy of a feature X and H(X|Y) the entropy of a feature X given the occurrence of feature Y the correlation between two features X and Y can then be calculated using the symmetrical uncertainty:

$$C(X|Y) = \frac{H(X) - H(X|Y)}{H(Y)} \quad (3)$$

In this algorithm, target class of a data sample is also considered to be a feature. In this research work, Best First search method has been used for subset search in the forward and backward directions [12], [15].

## 3.7 Consistency based Feature Selection Algorithm

Consistency based Feature Selection Algorithm, [16] is used to evaluate the subsets of features simultaneously and selects the optimal subset. This optimal subset is the smallest subset of features that can identify samples of a class as consistently as the complete feature set. In order to determine the consistency of a subset, the combination of feature values representing a class are given a pattern label. All samples of a given pattern should thus represent the same class. If two samples of the same pattern represent different classes, then that pattern is considered to be inconsistent. This algorithm gives subset of features which have very little number of features used to characterize each application class.

In this research work, Best First search method has been employed for subset search in the forward and backward directions.

## 3.8 Principal Components Analysis based Feature Selection Algorithm

Principal Components Analysis (PCA) based Feature Selection Algorithm, [17] is the predominant linear dimensionality reduction algorithms, which is widely applied on datasets in all scientific domains. PCA maps the data points from a high dimensional space to a low dimensional space while keeping all the relevant linear structure unchanged.

PCA is an unsupervised feature reduction technique in which only input samples are the coordinates of the data points and the number of dimensions.

This algorithm performs a principal components analysis and transformation of the data features. In this technique Ranker search method is used for subset search. The process of dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data (default percentage = 95%). Feature noise can be filtered by transforming to the Principal Components Space, eliminating some of the worst eigenvectors, and then transforming back to the original space.

# 4. INTERNET TRAFFIC DATASET

In this research work, Wireshark, [19], which is well-known packet capturing software, is used to capture real time internet traffic. It is a network packet analyzer which is used to capture network packets and displays fully detailed packet data. Real time internet traffic packets are captured for the duration of 2 minutes for each internet application just by considering on-going middle session of each application. During this packet capturing process, starting and end of each application are not taken into account.

In this process of developing datasets, two types of datasets are obtained: one is full feature dataset which is named as Dataset I and another are reduced feature datasets which are named as Dataset $II_{Correlation}$, Dataset $II_{Consistency}$ and Dataset $II_{PCA}$ [6]. In both types of datasets, seven internet applications are taken into account such as www, e-mail, web media, P2P, FTP data, instant messaging and VoIP. These datasets include 2800 samples.

In Dataset I, each internet application is characterized by 261 features which mainly consist of minimum, maximum, mean, variance and total values of no. of packets, average packets per second, packet size, duration, no. of conversations etc for Ethernet, IPv4, IPv6, TCP and UDP packets.

While reduced feature datasets are obtained from full feature dataset I using three feature selection algorithms of Weka Toolkit [12]. Dataset $II_{Correlation}$ is developed using Correlation based Feature Selection Algorithm and in this dataset, each internet application is characterized by 41 features. Dataset $II_{consistency}$ has been obtained using Consistency based Feature Selection Algorithm and in this dataset, each internet application is characterized by 9 features only. Third Dataset $II_{PCA}$ is developed using Principal Components Analysis based Feature Selection Algorithm and in this dataset, each data sample is characterized by 37 features.

For our work, we have used 2.27 GHz Intel core i3 CPU workstation with 3GB of RAM and Microsoft Windows 7 operating system.

# 5. IMPLEMENTATION AND RESULT ANALYSIS

## 5.1 Methodology

In this research work, Weka toolkit, [11] which is a well-known data mining tool, is used for implementing IP traffic

classification with five different ML algorithms. Four types different internet traffic datasets, Dataset I, Dataset $II_{Correlation}$, Dataset $II_{consistency}$ and Dataset $II_{PCA}$ consisting of 2800 data samples in each, are divided into two sets consisting of 2500 data samples for training and 300 data samples for testing purpose for all datasets.

In this work, classification accuracy, training time, recall and precision values [1], [4] of individual internet application samples are employed in order to evaluate performance of these five ML algorithms/classifiers. All these parameters are defined as follows:

- Classification Accuracy: It is the percentage of correctly classified samples over all classified samples.
- Training Time: It is the total time taken for training of a machine learning classifier. In this paper, it is measured in seconds.
- Recall: It is the proportion of samples of a particular class Z correctly classified as belonging to that class Z. It is equivalent to True Positive Rate (TPR). In this paper, its value ranges from 0 to 1.
- Precision: It is the proportion of the samples which truly have class z among all those which were classified as class z. In paper its value ranges from 0 to 1.

## 5.2 Results and Analysis

Table I shows classification accuracy and training time of MLP, RBF, C4.5 Bayes Net and Naïve Bayes ML classifiers for Dataset I. It is clear from this table and figure 4 that maximum classification accuracy is provided by Bayes Net classifier for full feature dataset which is 85.33 %. From table I, it is evident that training time of Bayes Net classifier is 14 second which is much lesser as compared to that of MLP and RBF classifiers in case of full feature dataset. But it is slightly larger than that of C4.5 and Naïve Bayes classifiers.

Table 1. Classification Accuracy and Training Time of five ML classifiers for Dataset I

| ML Classifiers | MLP | RBF | C4.5 | Bayes Net | Naïve Bayes |
|---|---|---|---|---|---|
| **Classification Accuracy (%)** | 31.33 | 82.33 | 79 | 85.33 | 68 |
| **Training Time (Seconds)** | 220 | 126 | 12 | 14 | 4 |

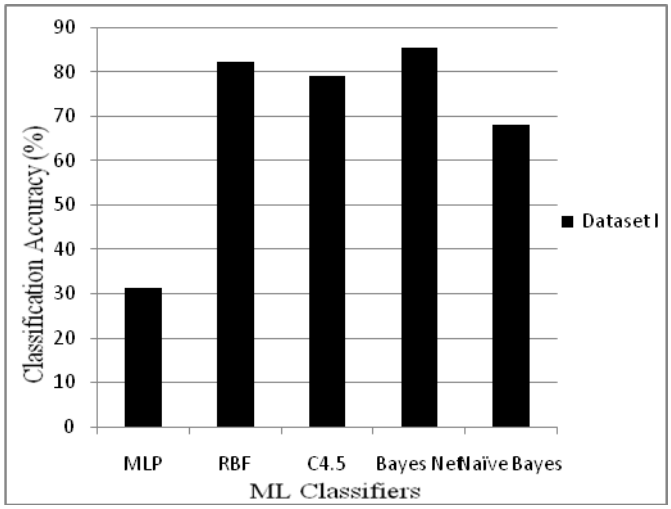


Fig 4: Classification Accuracy of five ML Classifiers for Dataset I

From these results, it is evident that Bayes Net gives better performance in terms of classification accuracy as compared to other four ML classifiers for Dataset I. Figure 5 and 6 shows recall and precision values of three most accurate ML classifiers i.e. RBF, C4.5 and Bayes Net for individual internet applications. Bayes Net gives 100% recall value for P2P, FTP data, instant messaging and VoIP applications. Similarly, it gives 100 % precision for Web media, P2P, IM and VoIP applications. Thus it is again clear that Bayes Net gives better performance in terms of Recall and precision for most of internet applications in case of full feature dataset I.

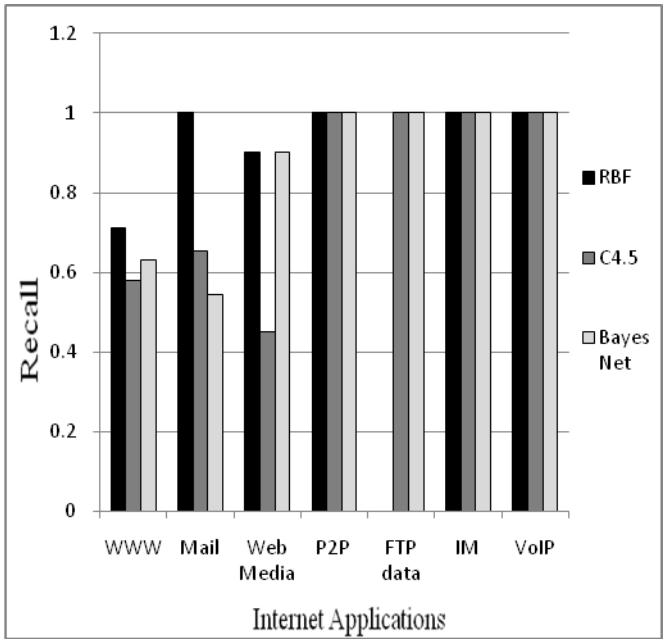


Fig 5: Recall of three most accurate ML Classifiers for Dataset I

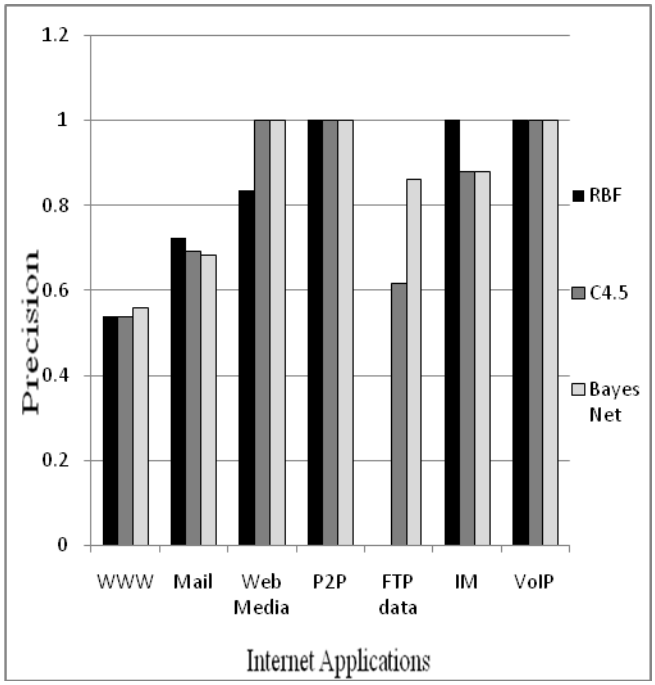


Fig 6: Precision of three most accurate ML Classifiers for Dataset I

Although, Bayes Net provide better classification performance at full feature dataset I. But there is main problem of large training time and complexity of this ML classifier which make this technique ineffective for online and real time IP traffic classification. There is still scope of further improvement in classification accuracy and reduction in training time and computational complexity if number of features used to characterize each internet application can be reduced to great extent. In the following section, three different feature selection algorithms are employed to reduce the number of features characterizing each internet applications to make three different reduced feature datasets and the performance of five ML algorithms has been evaluated on the basis of these reduced feature datasets.

Table 2 shows classification accuracy and training time of five ML algorithms for Dataset $II_{correlation}$ which is obtained using Correlation based Feature Selection Algorithm. It is evident from this table that highest classification accuracy is provided by C 4.5 algorithm with reduction in training time to 1 second. Using this dataset, Bayes Net also performs well for IP traffic classification with accuracy of 90% and training time of 2 seconds only.

Table 2. Classification Accuracy and Training Time of five ML classifiers for Dataset $II_{correlation}$

| ML Classifiers | MLP | RBF | C4.5 | Bayes Net | Naïve Bayes |
|---|---|---|---|---|---|
| **Classification Accuracy (%)** | 43 | 77 | 93.66 | 90 | 74.66 |
| **Training Time (Seconds)** | 56 | 89 | 1 | 2 | 1 |

Table 3 shows classification accuracy and training time of these ML algorithms for Dataset $II_{consistency}$ which is obtained from full feature dataset I using Consistency based Feature Selection Algorithm. In this case, again C 4.5 gives satisfactory performance in terms of classification accuracy of 84.33% and training time of 1 second only. But performance of ML algorithms is poor in case of Consistency based Feature Selection Algorithm as compared to that of Correlation based Feature Selection Algorithm.

Table 3. Classification Accuracy and Training Time of five ML classifiers for Dataset $II_{consistency}$

| ML Classifiers | MLP | RBF | C4.5 | Bayes Net | Naïve Bayes |
|---|---|---|---|---|---|
| **Classification Accuracy (%)** | 52.66 | 71 | 84.33 | 64 | 64.3 |
| **Training Time (Seconds)** | 34 | 218 | 1 | 1 | 1 |

Table 4 gives the classification accuracy and training time of five ML algorithms for Dataset $II_{PCA}$ which is obtained using Principal Components Analysis (PCA) based Feature Selection Algorithm. In this case, once again C 4.5 algorithm gives better classification performance in terms of classification accuracy of 87 % and training time of 2 seconds only. This overall performance in case of PCA based Feature Selection Algorithm is better is better than that of Consistency based Algorithm. But it is inferior to that of Correlation based Feature Selection Algorithm.

Table 4. Classification Accuracy and Training Time of five ML classifiers for Dataset $II_{PCA}$

| ML Classifiers | MLP | RBF | C4.5 | Bayes Net | Naïve Bayes |
|---|---|---|---|---|---|
| **Classification Accuracy (%)** | 48.33 | 80.66 | 87 | 49.33 | 72 |
| **Training Time (Seconds)** | 51 | 88 | 2 | 3 | 1 |

A bar graph in figure 7 is presented to visualize the performance of five ML algorithms and three feature selection algorithms. It is evident from this bar graph that maximum classification accuracy is obtained by C 4.5 algorithm when Dataset $II_{correlation}$ is taken into account. This reduced feature dataset is obtained from full feature dataset I using Correlation based Feature Selection Algorithm. Thus for feature reduction, Correlation based Feature Selection Algorithm performs well as compared to Consistency based and PCA based Feature Selection Algorithms. This combination of C 4.5 Decision Tree Algorithm with Correlation based Feature Selection Algorithm is very effective for IP traffic classification.

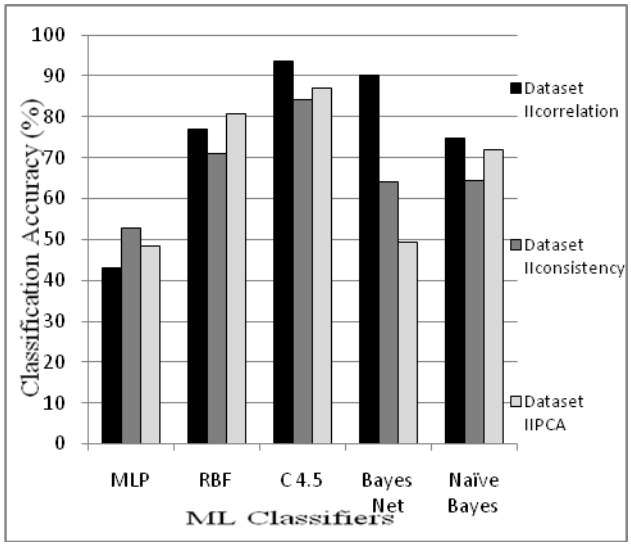


Fig 7: Classification Accuracy of five ML algorithms in case of three different reduced feature datasets.

From these results, it is clear that C4.5 gives better performance in terms of classification accuracy and training time as compared to other four ML classifiers when Correlation based Feature Selection Algorithm is taken into account for feature reduction process. Figure 8 and 9 shows recall and precision values of three most accurate ML classifiers i.e. RBF, C4.5 and Bayes Net for individual internet applications in case of Dataset $II_{Correlation}$. C4.5 gives 100% recall value for P2P, FTP data, instant messaging and VoIP applications. Similarly, it gives 100 % precision for Web media, P2P and VoIP applications. Thus It is again very clear that C4.5 gives better performance in terms of Recall and precision for most of internet applications in case of reduced feature dataset $II_{Correlation}$.

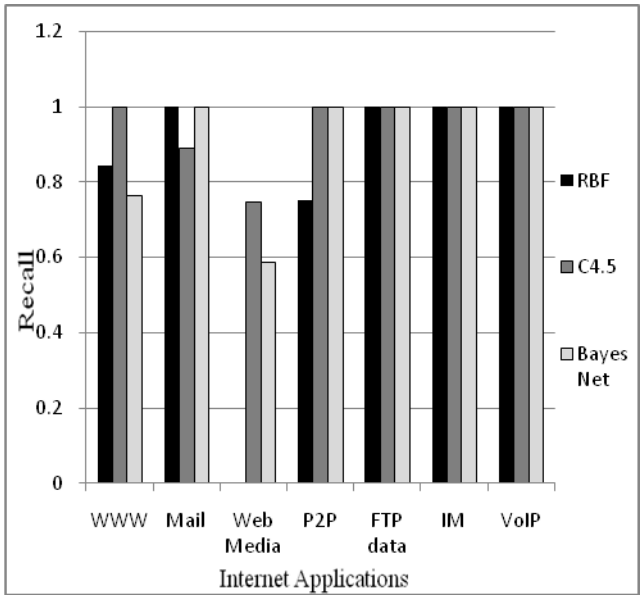


Fig 8: Recall of three most accurate ML Classifiers for Dataset $II_{Correlation}$

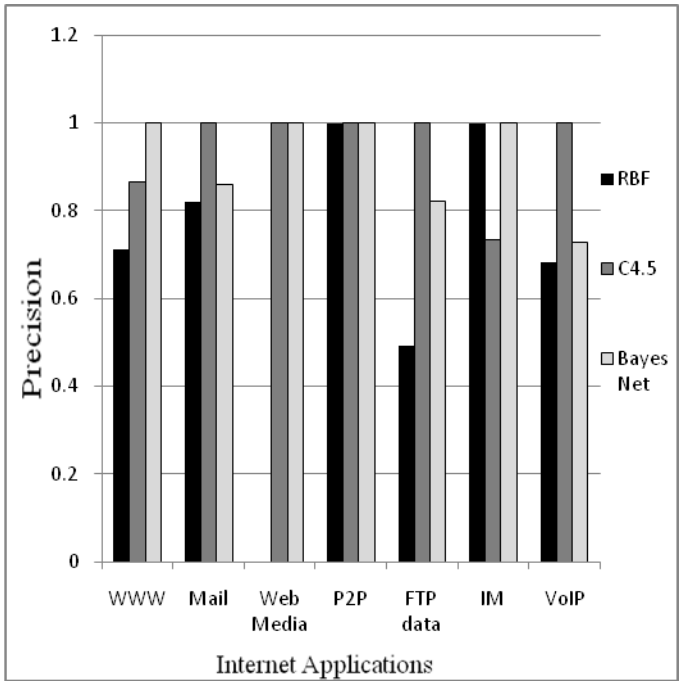


Fig 9: Precision of three most accurate ML Classifiers for Dataset $II_{Correlation}$

From this analysis, it is evident that Bayes Net gives good performance in terms of classification accuracy, training tine, recall and precision of individual internet applications for full feature dataset I. But this performance is still not remarkable. So by reducing the number of features used to characterize internet applications, classification performance is further improved. For feature reduction process, Correlation based Feature Selection Algorithm performs well for IP traffic classification as compared to other two algorithms. Thus using Correlation based feature Selection Algorithm for feature reduction, C4.5 gives highest degree of performance in terms of all the factors mentioned above for IP traffic classification.

## 6. CONCLUSIONS AND FUTURE SCOPE

In this paper, firstly real time IP traffic has been captured using Wireshark software which is a packet capturing tool. After that, Internet traffic is classified using five ML classifiers. Results show that Bayes Net gives better classification of internet traffic data in terms of classification accuracy, training time of classifiers, recall and precision values of classifiers for individual internet applications. But the main problem is of training time and computational complexity of Bayes Net Algorithm which make this classification technique effective only for offline IP traffic classification. In order to make ML techniques effective for online and real time IP traffic classification, the number of features used to characterize each internet application data sample of this full feature dataset are further reduced using three different Feature Selection Algorithms to make a reduced feature datasets. Results show that with reduced feature dataset obtained by using Correlation based Feature Selection Algorithm, C 4.5 algorithm gives better performance in terms of classification accuracy and training time. Thus, it is evident that C4.5 is an efficient ML technique for IP traffic classification with reduction in number of features

characterizing each internet application using Correlation based Feature Selection algorithm.

In this research work, internet traffic dataset has been developed by considering packet capture duration of 2 minutes for each application which is still very large. This packet capture duration can be further reduced to make evaluation more real time compatible. Secondly, internet traffic can also be captured from various different real time environments such as university or college campus, offices, home environments etc. This internet traffic dataset can be extended for many other internet applications which users use in their day to day life.

## 7. REFRENCES

[1] Thuy T.T. Nguyen and Grenville Armitage. "A Survey of Techniques for Internet Traffic Classification using Machine Learning," IEEE Communications Survey & tutorials, Vol. 10, No. 4, pp. 56-76, Fourth Quarter 2008.

[2] Arthur Callado, Carlos Kamienski, Géza Szabó, Balázs Péter Ger˝o, Judith Kelner,Stênio Fernandes ,and Djamel Sadok. "A Survey on Internet Traffic Identification," IEEE Communications Survey & tutorials, Vol. 11, No. 3, pp. 37-52, Third Quarter 2009.

[3] Runyuan Sun, Bo Yang, Lizhi Peng, Zhenxiang Chen, Lei Zhang, and Shan Jing. "Traffic Classification Using Probabilistic Neural Network," in Sixth International Conference on Natural Computation (ICNC 2010), 2010, pp. 1914-1919.

[4] Kuldeep Singh and Sunil Agrawal, "Internet Traffic Classification using RBF Neural Network," in International Conference on Communication and Computing technologies(ICCCT-2011), Jalandhar, India, February 25-26, 2011, paper 10, p.39-43.

[5] Andrew W. Moore and Denis Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," in *SIGMETRICS'05*, Banff, Alberta, Canada , June 6.10, 2005.

[6] Luca Salgarelli, Francesco Gringoli, Thomas Karagiannis. "Comparing Traffic Classifiers," ACM SIGCOMM Computer Communication Review, Vol. 37, No. 3, pp. 65-68, July 2007.

[7] Andrew W. Moore, Denis Zuev, Michael L. Crogan. 2005. Discriminators for use in flow-based classification. Queen Mary University of London, Department of Computer Science, RR-05-13, August 2005.

[8] Y.L. Chongand K. Sundaraj, "A Study of Back Propagation and Radial Basis Neural Networks on ECG signal classification," in 6th International Symposium on Mechatronics and its Applications (ISMA09), Sharjah, UAE, March 24-26, 2009.

[9] Mutasem khalil Alsmadi, Khairuddin Bin Omar, Shahrul Azman Noah ,Ibrahim Almarashdah, "Performance Comparison of Multi-layer Perceptron (Back Propagation, Delta Rule and Perceptron) algorithms in Neural Networks" in 2009 IEEE International Advance Computing Conference (IACC 2009) ,Patiala, India, 6-7 March 2009, p. 296-299.

[10] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees" Image Processing Division, National Institute for Space Research – INPE, SP, Brazil.

[11] Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2th edition, Morgan Kaufmann Publishers, San Francisco, CA, 2005.

[12] Weka website. Available: http:// www.cs.waikato.ac.nz/ml/weka/

[13] Jie Cheng, Russell Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System," Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada.

[14] Ioan Pop, "An approach of the Naive Bayes classifier for the document classification," General Mathematics, Vol. 14, No. 4, pp.135-138, 2006.

[15] Mark A. Hall. 1999. Correlation based Feature Selection for Machine Learning. University of Waikato, Hamilton, New Zealand, April, 1999.

[16] Manoranjan Dash, Huan Lau, "Consistency – based search in feature selection", Artificial Intelligence, Elsevier, 27 March, 2003.

[17] Christos Boutsidis, Michael W. Mahoney, Petros Drineas, "Unsupervised Feature Selection for Principal Components Analysis", KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

[18] Simon Haykin, Neural Networks: A Comprehensive foundation, 2th edition, Pearson Prentice Hall, New Delhi, 2005.

[19] Wireshark, Available: http:// www.wireshark.org/