Classifying Internet Traffic using An Efficient Classifier

Haitham A.Jamil, Hind G. Abdelrahim, Bushra M Ali, Azza O. Awad

Abstract— **The new development in the architecture of Internet has increased internet traffic. The introduction of Peer to Peer (P2P) applications are affecting the performance of traditional internet applications. Network optimization is used to monitor and manage the internet traffic and improve the performance of internet applications. The existing optimizations methods are not able to provide better management for networks. Machine learning (ML) is one of the familiar techniques to handle the internet traffic. It is used to identify and reduce the traffic. The lack of relevant datasets have reduced the performance of ML techniques in classification of internet traffic. The aim of the research is to develop a hybrid classifier to classify the internet traffic data and mitigate the traffic. The proposed method is deployed in the classification of traffic traces of University Technology Malaysia. The method has produced an accuracy of 98.3% with less computation time.**

Keywords*— *Classifier, Internet traffic, P2P, Internet traffic Mitigation, Machine Learning, Neural Network

I. INTRODUCTION

P2P application is one of the familiar developments in internet architecture[1]. Traffic data classification is used to reduce the internet traffic. The classified data will inspect the network and try to consume less bandwidth[2]. Managing network traffic is the major challenge in distributed network. Existing researches are not commendable in the area of P2P computing. The differences in the properties of traffic data will affect the performance of network. ML methods are used to develop an automated application to deal complex data. Dataset and feature selection are the important factors of classifiers. The ability of classifiers are dependent on these factors.

ML is a set of technique and data analytical methods to improve the performance of a automated task. ML algorithms are widely used in real – time applications[3].

Support Vector Machine, C5.0, and Artificial Neural Network (ANN) are some of the ML techniques used for the classification of P2P network. SVM is a supervised method that uses mapping function for the classification[4]. The mapping function is used to classify the data according to the labels. The classified model represents the data that was used in the training phase[5].

C5.0 is a familiar technique used as an alternative for SVM. It is based on Decision tree algorithm. The concept of estimation of entropy is used to take decsions[6][7]. The features of dataset will be used for the derivation of patterns and matched with a target class[8][9][10]. J48 algorithm is used in the selection of features by evaluating a node of the tree. It is a flexible method and easy to combine with other ML techniques[11][12][13].

Most of the complex ML models were built with NN and produced an optimal solution. It is basically a time consuming method[15][16]. A proper training is required for NN to produce good results.

The objective of the research is to provide a efficient Artificial NN (ANN) for the classification of P2P internet traffic. ANN is combined with J48 to improve the computation time of the classifier[17][18]. The proposed model can be used to identify the emergence of new P2P traffic application in distributed networks. The state of the art classifiers are compared with the proposed method in terms of accuracy and computation cost[18].

The structure of the paper is arranged as follows: section two will provide information about the existing literature on network traffic classification. Section three will give details about the methodology of the research. Section four will discuss the experimental setup of the research. Section five will provide results and analysis and finally, the paper will be concluded in section six.

II. REVIEW OF LITERATURE

Internet classification is used to provide better quality of service is managing network traffic. A classifier will cluster the internet traffic data into different groups. The new improvements such as dynamic port numbers an packet payloads are complex and difficult to classify using older classification technique.ML algorithms are need to be employed to classify complex communication data. Pramitha P et.al. ,(2018)[1] have compared ML algorithms for classification of internet traffic data.

The authors have employed Naïve Bayes, Random Forest(RF), Decision Tree(DT), and Multi layer perceptron algorithms for the classification of internet traffic. The results have shown that RF and DT have better accuracy than other classifiers.

Manuscript published on 30 September 2019
* Correspondence Author

Haitham A. Jamil, University of Elimam Elmahdi, Kosti, White Nile, Sudan
haithamjamil@mahdi.edu.sd

Bushra M Ali, University Technology Malaysia, Johor Bahru, Malaysia
bushra0912115@gmail.com

Hind G. Abdelrahim, University Technology Malaysia, Johor Bahru, Malaysia
hindjamil33@gmail.com

Azza O. Awad,University Technology Malaysia, Johor Bahru, Malaysia
azzaawadelkareem@gmail.com

Retrieval Number: B2041078219/19©BEIESP
DOI:10.35940/ijrte.B2041.098319

Wujian Ye and Kyungan Cho (2014)[2] have developed a hybrid classifier with heuristic rules. Signature and statistics based classifiers were developed by the authors for classification of unknown traffic.

Port based classification became void due to the nature of producing unreliable results. The introduction of new techniques in Internet technology has provided facility to hide port numbers. False positive and negative can be detected using this type of classification.

Payload based classifier will work beyond the transport layer to find unique values in packet payloads. Deep Packet Inspection(DPI) and Stochastic Payload Inspection (SPI) are the two categories of payload based classifiers. The payload of packets can be used to generate pattern of unusual traffic in Internet. DPI is effective, but the cost of computation is high. The method need to have continuous access of the packet memory. SPI have some efficient methods to compute unique pattern using protocols. Justin Ma et. al[3] have proposed a method to find a common string in packets, which is based on SPI.

Statistical classification is one of the effective methods in Internet traffic data classification. It will use flow level measurements for the characterization of different applications. ML techniques are easily implemented in this type of classifications.

Behavioral classification is a modern technique to characterize the network traffic. It has the capability to find applications running on the target host.

In [4], author have studied the abilities of different traffic classifiers. Authors[5] have proposed a two step hierarchical scheme for the detection of attacks on the web server. Palmieri F et.al.[6] , have found an approach to detect network anomalies on independent components. The study has used a ML technique for the detection of unusual patterns. Gil GD[7] have characterized virtual private network traffic using time related features. Authors [8] did a survey on different techniques in traffic classification. In [9],author have compared some supervised ML algorithms for the classification of network traffic.

The proposed research is based on behavioral classification and used a hybrid ML algorithm for the classification of P2P traffic and mitigates the traffic in network.

III. METHODOLOGY

The section will describe the process of classification of internet traffic data. The fig. – 1 will show the framework for classification of communication traffic.

A. Data Collection

Communication data are difficult to process by ML algorithms. SNORT is an old Intrusion detection system (IDS). The rules of SNORT are followed to match the data with signature in the database. It will label the data according to the application protocol. The logs are downloaded from a decoder as a Tcpdump format. SNORT can be replaced with modern IDS. The research has used SNORT due to its ability in finding signatures.

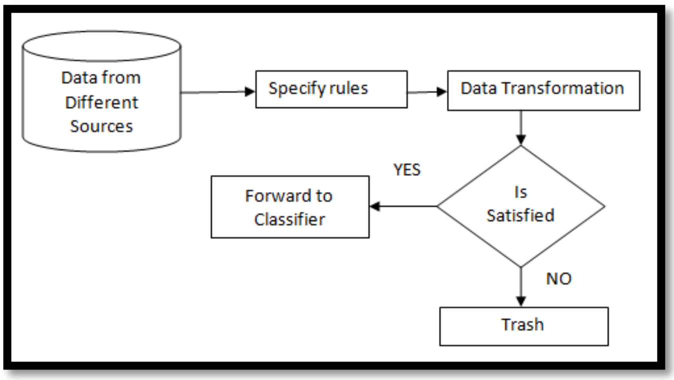

Fig. 1 Data collection and feature Extraction

B. Feature Selection

Feature selection or extraction is the process used in the pre – processing of data. It is used to extract optimal subset of features. ML algorithms will work on the features for the extraction of patterns. The consistency based feature selection is used in the work for the selection of features.

C. Behavioral based Classifier

Behavior based classifier will use the flow information of the network and classify the traffic. Packets will be parsed into tokens, easily accessible by ML algorithms. Fig. 2 is the illustration of offline classifier. The traffic traces will be pre – processed by a well defined pre – processor model. The features will be selected from the traces. SNORT rules will be used to find out the signatures from the traces. The concept of signature is used to filter the traces into features. The extracted features will be classified by the offline classifier. The offline classifier is used to train and test the results.

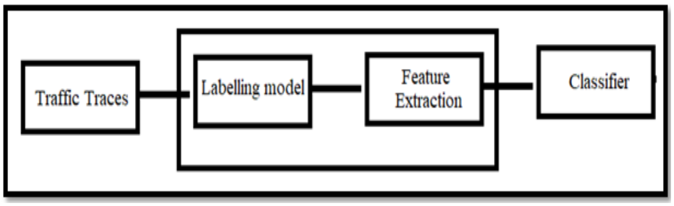

Fig. 2 Offline Classifier

Fig. 3 depicts an online classifier. An online classifier is connected to the network and traffic traces will be fed dynamically into the classifier. The classified traces will be used to mitigate the traffic in the network. The traffic will be mitigated dynamically without any human intervention.

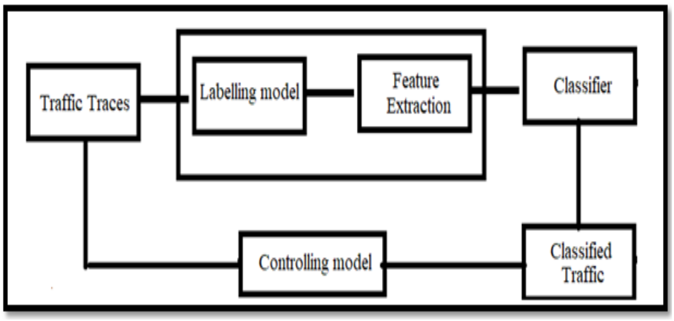

Fig.3 Online Classifier

IV. EXPERIMENTAL SETUP

The section will discuss the experimental setup of the proposed system. SNORT has to be configured to capture the flow from the online P2P network. It will further discuss the process to select flow features for fast computation

A. Capturing flow information

Collection of data from P2P will be performed using SNORT rules[19]. The new version 2.9.12 of SNORT is used in the research. Data will be stored in comma separated values(CSV) format. The configuration file is manipulated to capture the related signatures of P2P communication. The payload packets will be inspected by SNORT and related data will be stored according to the rules.

B. Selection of features

The focus of the research is to reduce the set of features and enhance the classification of traffic data set. Chi – Square and fuzzy rough algorithms were employed to extract features from dataset.

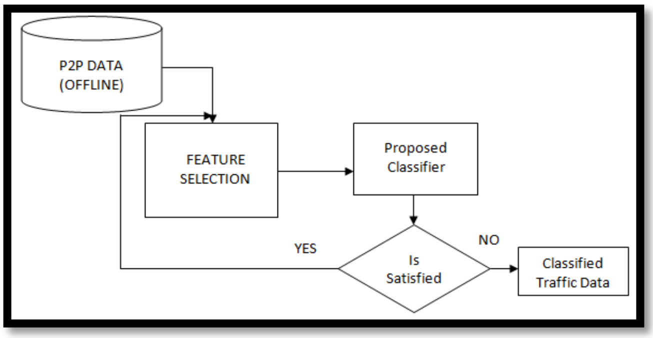

Fig.4 Feature selection and Offline Classification

Fig.4 will show the process of feature extraction and offline classification. Chi – square have the ability to select optimal features from the dataset. Online P2P ML algorithm
SVM, Random Forest(RF), ANN and C5. 0 are familiar algorithms in Traffic data classification. The following algorithms were selected after the careful study on familiarity and state of the art techniques from literature review.SVM, RF, ANN are the algorithms that are used with the proposed method. ANN is combined with J48 algorithm for the proposed work. ANN is slow learner. J48 is a decision based algorithm, which has the ability to increase the speed of ANN. Fig. 5 shows the online P2P classifier. The classified data will be fed into the network and reduce the traffic. The source of the data will be updated each time with data from the real time network.

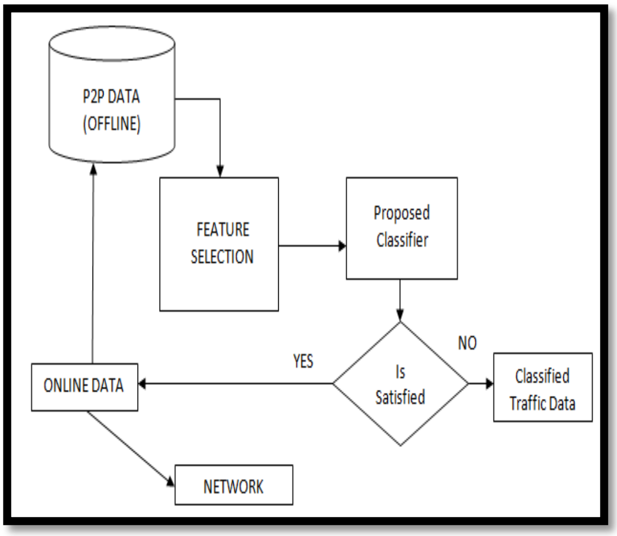

Fig.5 Online P2P classifier

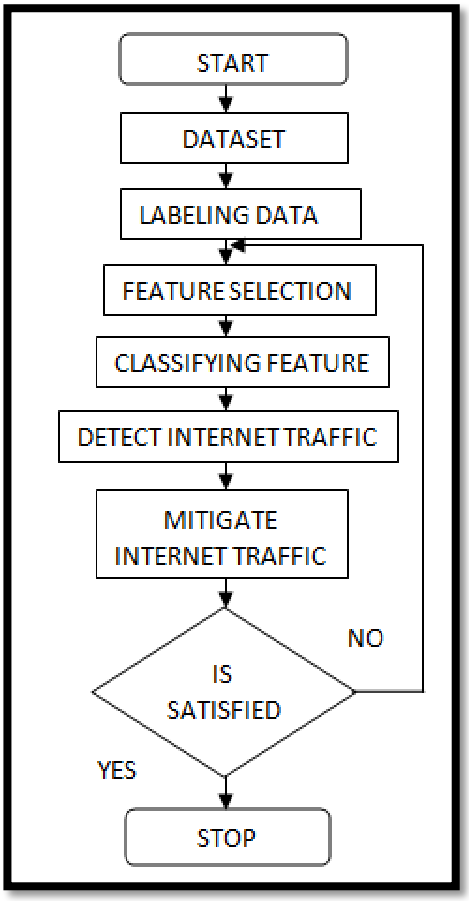

Fig. 6 Flow of Process – P2P Classifier

Fig. 6 shows the processes involved in the classification of data. Labeling data is the important process for the selection of features. The following Algorithm 1 shows the classification algorithm followed by the research.

```
Algorithm1 P2P classifier
Input:
       D: { D1,D2,.. .. ,Dn}-> Dataset
       Ft: { Ft1, Ft2,.. .., Ftn} -> Feature Extraction
Output:
       { Classified Traffic}
Algorithm:
Data: {Dataset}
LData: for each Data in Dataset
            {  Label Data using Snort rules}
For each file in packet
   If IP_source && IP_dest belong to flow
     Add Packet to dataset
  Else
    New Packet Generation
Read Dataset D1,D2…,Dn
Split Data into tokens
Feature Extraction;
Start Classifier
Input No. of Neurons
Input No. of Hidden Layers
Build Model
Classify traffic
Stop Classifier
Input Classifier Traffic in Network
End
```

The classifier will return the results to the controlling model to update the new classified data into the real time network.

V. RESULTS AND DISCUSSION

The section will give results and evidences to prove the efficiency of proposed method. Datasets were collected from different resources. Global evaluation metrics were applied to evaluate the algorithms.

Table I is the details of dataset. University Technology Malaysia(UTM) datasets were collected in three different period. The period between July – October 2011, three datasets were collected from the network. The second period was October 2012 and final dataset were downloaded in November 2012.

Table I –Details of datasets

Source	Datasets	Average Packets	Average Flows	Average Size (MB)
UTM	5	138456	8423	542
CAIDA	1	88564	2540	845
UNIBS	1	8457856	78998	1345
Cambridge	1	1145823	45868	878

CAIDA[20] datasets are available for researchers without any cost. Datasets are collected in the period between 2009 and 2013. University of Brescia (UNIBS)[21] had offered the traces collected between September and October 2009. Cambridge[22] datasets were collected in the month August 2003. Datasets were old in nature, but flow information will be useful to train the algorizhms to produce better results.

Precision and Recall are used to evaluate the retrieval capacity of the method.True positive (TP) and False Positive(FP) are the important metrics used to prove the quality of classification of ML algorithms. Finally, accuracy is calculated and compared with each other algorithms.

Methods	Building Time (Seconds)				
	FS1	FS2	FS3	FS4	FS5
SVM	2.6	0.25	0.46	1.78	0.86
RF	5.3	3.45	3.68	5.4	4.3
ANN	12.56	5.36	7.37	6.3	7.8
Proposed Method	1.45	0.48	1.3	2.4	1.89

Table II Details of Building Time

Table III TP and FP (Training Time)

Classifier	TP	FP
SVM	96.8	3.2
RF	98.2	1.8
ANN	97.6	2.4
Proposed Method	98.4	1.6

Table IV TP and FP (Testing Time)

Classifier	TP	FP
SVM	**98.3**	**1.7**
RF	**97.6**	**2.4**
ANN	**98.7**	**1.3**
Proposed Method	**98.9**	**1.1**

Table II provide details of the building time of the models. The proposed method and SVM took less time comparing to other models. SVM is more efficient to produce results in less amount of time. The proposed ANN is a feed forward network that have auto heal property. J48 is also a factor to have less building time for proposed method.

Tale III and IV are showing details of TP and FP of methods applied in the research. All ML algorithms have produced better FP and TP because of its patten recognition property. The performance of ML in classificaiton and clustering problems are better than other algorithms.

Table V Details of Precision and Recall

Methods	FS1		FS2		FS3		FS4		FS5	
	P	R	P	R	P	R	P	R	P	R
SVM	0.67	0.72	0.81	0.78	0.84	0.81	0.78	0.74	0.81	0.84
RF	0.85	0.82	0.84	0.91	0.78	0.86	0.94	0.91	0.91	0.87
ANN	0.81	0.78	0.82	0.86	0.91	0.87	0.79	0.86	0.84	0.86
Proposed Method	0.79	0.8	0.87	0.91	0.89	0.91	0.93	0.92	0.89	0.86

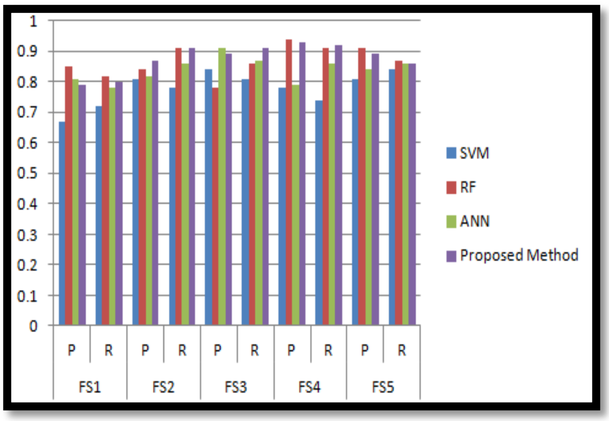

Fig. 7 precision and Recall

Table V and fig. 7 shows the precision and recall of the methods. The proposed method have better values than other methods. The performance of RF,ANN and proposed methods are better compairng to SVM. Basically, SVM does not have better precision, recall and F1 score.

Table VI provides details of accuracy of the methods. Fig. 8 has shown the results related to table VI.

Table VI Accuracy of Classifiers

Methods	Accuracy			
	DS – 1 (%)	DS – 2 (%)	DS –3 (%)	DS – 4 (%)
SVM	89.1	91.3	88.6	90.5
RF	91.3	92.6	94.4	94.6
ANN	97.3	96.8	95.3	92.3
Proposed Method	96.8	98.2	96.3	93.6

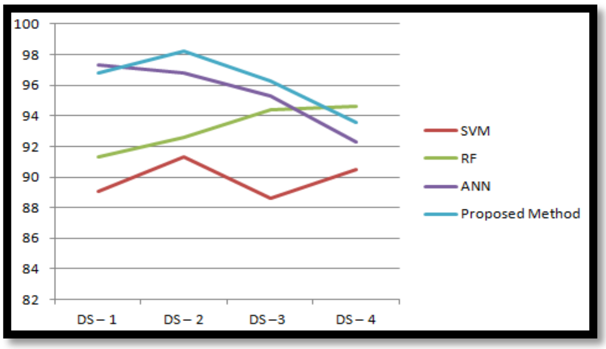

Fig. 8 Classifiers Accuracy

Both ANN and proposed method have better accuracy. The ANN model will take more time for training, but accuracy will be more. Proposed method has better overall performance comparing to other methods.

CONCLUSION

The introduction of modern technology in the world of internet has raised the data traffic in network. P2P communication needs reliable traffic to share messages. The proposed study has combined both ANN and J48 algorithms for the classification of internet traffic data. The overall performance of the proposed method is better comparing to other methods. The proposed method has 98.9% of True positive and an average accuracy of 96. 2%. It is evidence that the proposed method is more reliable than other methods

ACKNOWLEDGMENT

The authors are thankful to University of Elimam Elmahdi for rendering support and guidance. We like to thank University of Malaysia for offering dataset.

REFERENCES

1. Perera P, Tian Y C, Fidge C, et al (2017) A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic,International Conference on Neural Information Processing. Springer, Cham, 445-454
2. Ye W, Cho K (2014) Hybrid P2P traffic classification with heuristic rules and machine learning[J]. Soft Comput 18(9):1815–1827
3. J. Ma, K. Levchenko, C. Kreibich, S. Savage, and G. M. Voelker,“Unexpected means of protocol inference,” in Proceedings of the 6thACM SIGCOMM conference on Internet measurement. ACM, 2006,pp. 313–326
4. Zhang J, Xiang Y et al (2013) Network traffic classification using correlation information. IEEE Trans Parallel Distrib Syst 24(1):104–117
5. Choi B, Cho K (2013) Two-step hierarchical scheme for detecting detoured attacks to the web server. Comput Sci Inf Syst 10(2):633–649
6. Palmieri F, Fiore U, Castiglione A (2013) A distributed approach to network anomaly detection based on independent component analysis. In: Concurrency and computation: practice and experience.
7. Finsterbusch M, Richter C, Rocha E, Muller JA, Hanssgen K (2014) A survey of payload-based traffic classification approaches. IEEE Communications Surveys & Tutorials 16 (2):1135–1156
8. Gil GD, Lashkari AH, Mamun M, Ghorbani AA, (2016) Characterization of encrypted and vpn traffic using time-related features. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016), pp 407–414.
9. Perera P, Tian YC, Fidge C, Kelly W (2017) A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. In: Neural Information Processing, Springer, Cham, Lecture Notes in Computer Science, pp 445– 454, DOI 10.1007/978-3-319-70087-8 47

10. J. Zhang, C. Chen, Y. Xiang, W. Zhou, Y. Xiang, Internet traffic classification by aggregating correlated naive bayes predictions, IEEE Trans. Inform. Forens. Secur. 8 (1) (2013) 5–15
11. L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, K. Salamatian, Traffic classification on the fly, ACM SIGCOMM Comput. Commun. Rev. 36 (2) (2006) 23–26.
12. G. Gómez Sena, P. Belzarena, Early traffic classification using support vector machines, in: Proceedings of the 5th International Latin American Networking Conference, ACM, 2009, pp. 60–66
13. Justin Ma, Kirill Levchenko, Christian Kreibich, Stefan Savage, and Geoffrey M. Voelker. Unexpected means of protocol inference. In 6th ACM SIGCOMM Internet Measurement Conference (IMC'06), Rio de Janeiro, BR, October 2006
14. Silvio Valenti, Dario Rossi, Alberto Dainotti, Antonio Pescape, Alessandro Finamore and Marco Mellia, " Reviewing Traffic Classification", - Reviewing Traffic Classification - Data Traffic Monitoring and Analysis: From Measurement, Classification, and Anomaly Detection to Quality of Experience,Part of the Lecture Notes in Computer Science book series (LNCS, volume 7754)
15. H. A. Jamil, A. M, A. Hamza, S. M. Nor, and M. N. Marsono, "Selection of online Features for Peer-to-Peer Network Traffic Classification," in Recent Advances in Intelligent Informatics. vol. 235, ed: Springer International Publishing, 2014, pp. 379-390.
16. M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh,W. Lee, and D. Dagon, "From throw-away traffic to bots: detecting therise of dga-based malware," in Presented as part of the 21st USENIXSecurity Symposium (USENIX Security 12), 2012, pp. 491–506.A. W. Moore and D. Zuev, "Internet traffic classification using bayesiananalysis techniques," in ACM SIGMETRICS Performance EvaluationReview, vol. 33, no. 1. ACM, 2005, pp. 50–60.
17. H. Dahmouni, S. Vaton, and D. Rossé, "A markovian signature-basedapproach to ip traffic classification," in Proceedings of the 3rd annualACM workshop on Mining network data. ACM, 2007, pp. 29–34.
18. H. L. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Z. Yu, "Feature selection for optimizing traffic classification," Computer Communications, vol. 35, pp. 1457-1471, Jul 1 2012.
19. SNORT Network Intrusion Detection System. Available: www.snort.org
20. The Cooperative Association for Internet Data Analysis. Available: http://www.caida.org/data
21. (19 Nov). Università Brescia data sets. Available: http://www.ing.unibs.it/ntw/tools/traces/download/
22. Cambridge data sets. Available: http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/index.html

AUTHORS PROFILE

Haitham Ahmed Jamil Mohammed is assistant professor at Elimam Elmahdi university. He received the B.Sc. and M.Sc. from University of Gezira Sudan and PhD , from Universiti Teknologi Malaysia. His research interests include computer network, Network Traffic classification, Peer-to-Peer computing, and optimization techniques.

Bushra Mohammed Ali Abdalla received the B.Sc. and M.Sc. from University of Gezira, Sudan. He is a lecturer at the faculty of Computer and Statistics Studies, University of Kordofan. Currently he is a PhD candidate at the School of Electrical Engineering, Universiti Teknologi Malaysia.

Hind Gamil Abdelrahim Gamil is currently pursuing her Ph.D. in Electrical Engineering at School of Electrical Engineering, Universiti Teknologi Malaysia. Her research interests include wireless sensor network, software defined network and network optimization techniques

Azza Omer Awad Elkreem is currently doing her Ph.D. in Electrical Engineering at School of Electrical Engineering, Universiti Teknologi Malaysia. Her research interests include communication Network, network-on-chip, and optimization techniques.