ISSN No:-2456-2165

Network Traffic Classification Techniques and Comparative Evaluation of Machine Learning Models

Mohsin Mahmood, Sohaib Ahmad Khalil, Syed Jamil Shah, Mansoor Ahmad
National University of Computer & Emerging Sciences, Pakistan

Abstract:- This paper stresses the importance of network traffic classification for ISPs in managing network applications effectively. It provides a step-by-step process for classification, using a real-time dataset and four machine learning classifiers. The C4.5 classifier achieves the highest accuracy among the tested classifiers. Network traffic classification optimizes performance, improves customer service, and enables the detection of security threats. Machine learning, particularly the C4.5 classifier, is effective in achieving accurate results. Overall, network traffic classification is crucial for ISPs, and machine learning techniques enhance network performance and customer experience.

***Keywords**:- Traffic Classification; Machine Learning; Methods.*

I. INTRODUCTION

Network Traffic Classification is an increasingly important field within Computer Science. It plays a crucial role in the overall management and performance of networks for Internet Service Providers (ISPs). The first step in network traffic classification is identifying and categorizing unknown network classes. This process is vital for network security and management, including tasks like Intrusion Detection and Quality of Service (QoS). By employing traffic classification techniques, network operators can make informed decisions such as blocking specific flows and efficiently managing network resources. Additionally, they can monitor the growth of network applications. Over the past two decades, various network traffic classification techniques have been proposed. The Port-Based Technique, which relies on registered port numbers, was initially effective but faced challenges due to the rise of Peer-to-Peer (P2P) applications that utilize dynamic port numbers. The Payload-Based Technique, also known as Deep Packet Inspection (DPI), offers accurate results but falls short when it comes to encrypted data network applications, as many applications employ encryption to protect data from detection. To overcome these limitations, researchers have introduced Machine Learning Techniques (ML) for internet traffic classification. ML techniques have shown promising results in network traffic classification by leveraging training and testing datasets to classify unknown classes effectively. This approach enables the identification of different types of applications flowing through the network with high accuracy.

- *Contribution*

In this paper, we delve into network traffic classification techniques and provide a comparative analysis of four machine learning classifiers. To begin, we capture network traffic using the packet capturing application Wireshark [15]. Next, we utilize the NetMate tool [21] to extract features from the captured traffic. Subsequently, we employ four machine learning classifiers to classify applications such as WWW, DNS, FTP, P2P, and Telnet. Based on experimental results, the C4.5 classifier demonstrates the highest accuracy among the machine learning classifiers, achieving an accuracy rate of 78.91%.The remaining sections of the paper are structured as follows: Section II presents introductory information about the developed techniques. Section III outlines the Internet Traffic Classification Model. Finally, in Section IV, we provide a conclusion based on our findings.

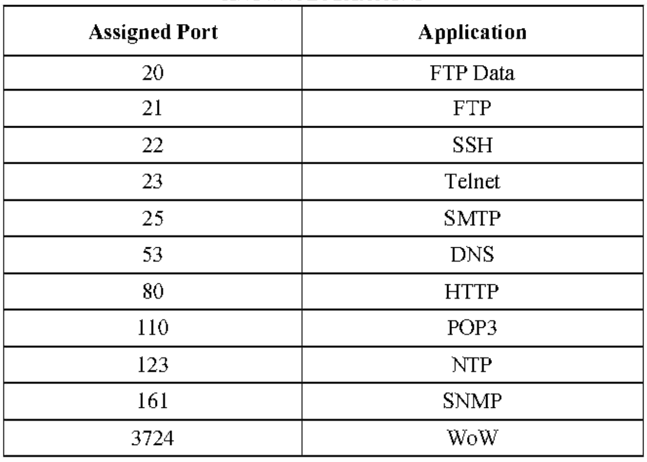
II. NETWORK TRAFFIC CLASSIFICATION TECHNIQUES

Network Traffic Classification involves identifying network applications or protocols in a network. It has become increasingly important in the past 20 years, and various methods have been proposed by researchers, including Port-based, Payload-based, and Machine Learning techniques. Port-based technique involves examining the port number of network traffic, while Payload-based technique examines the content of the data. Machine Learning techniques involve using algorithms to learn from data and classify network traffic accordingly.

A. Port-Based Technique

As discussed in Section I, the traditional approach to network application classification relies on utilizing well-known port numbers. However, this technique faces certain challenges. Firstly, it requires expensive hardware capable of pattern searching within the payload. Secondly, it is ineffective in classifying encrypted network application traffic. Lastly, this approach necessitates regular updates to the signature patterns of new applications to maintain accurate classification.

TABLE I. IANA ASSIGNED PORT NUMBER FORMAT FOR SOME WELL-KNOWN APPLICATIONS

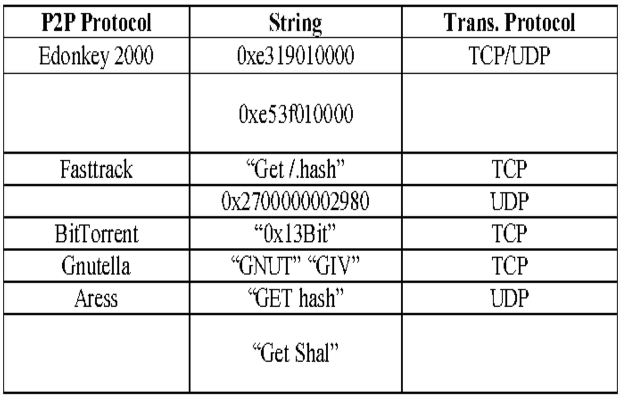


Assigned Port	Application
20	FTP Data
21	FTP
22	SSH
23	Telnet
25	SMTP
53	DNS
80	HTTP
110	POP3
123	NTP
161	SNMP
3724	WoW

B. *Payload-Based Technique*

The Deep Packet Inspection (DPI) technique examines packet contents for network application signatures, providing an alternative to the ports-based method. Specifically designed for Peer-to-Peer (P2P) applications that use dynamic port numbers to identify traffic, DPI is the first alternative to the ports-based method. Karayiannis et al. provide examples in [5]. However, DPI faces challenges, such as requiring a high output of machine learning, and the inability to classify instances into predefined classes. In this technique, machine learning training does not identify or classify instances in predefined classes, which is the first problem encountered. Despite the benefits of DPI, there are significant limitations to its effectiveness, which must be addressed in future research.

TABLE II. KARAGIANNIS DESCRIBE STRINGS AT THE BEGINNING P2P PROTOCOL PAYLOAD



P2P Protocol	String	Trans. Protocol
Edonkey 2000	0xe319010000	TCP/UDP
	0xe53f010000	
Fasttrack	"Get /.hash"	TCP
	0x2700000002980	UDP
BitTorrent	"0x13Bit"	TCP
Gnutella	"GNUT" "GIV"	TCP
Aress	"GET hash"	UDP
	"Get Shal"	

C. *Machine Learning (ML) Technique*

Machine learning techniques for network traffic classification utilize labeled data sets to train classifiers, which can then classify unknown traffic. Supervised learning involves training the classifier with labeled data, while unsupervised learning enables the classifier to identify patterns in unlabeled data. These methods have demonstrated positive outcomes in enhancing the precision and effectiveness of traffic classification tasks.

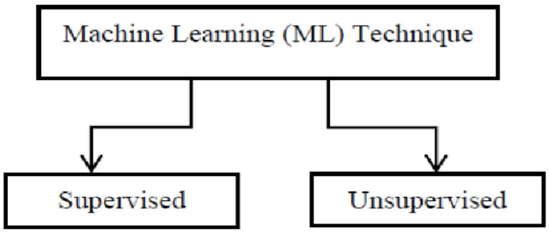


Figure 1. Kinds of machine learning

D. *Supervised Learning Technique*

Supervised learning is a type of machine learning technique used for classification tasks. It requires a complete labeled data set to train the model, which can then classify unknown classes. Supervised learning involves training a model using input-output pairs, where the input is the data and the output is the desired result. Once the model is trained, it can make predictions on new data samples. Supervised learning has been widely used in traffic classification due to its ability to handle complex data sets and provide accurate results.

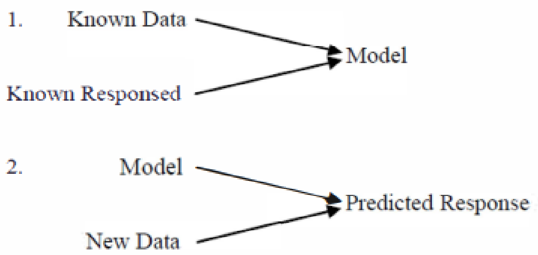
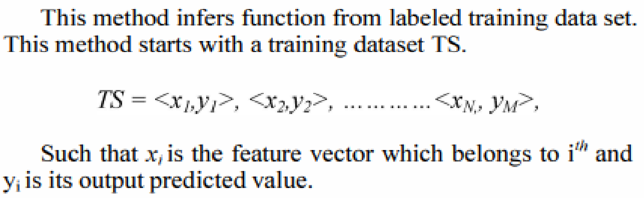


Figure 2. Method description by mathworks



This method infers function from labeled training data set. This method starts with a training dataset TS.

$$TS = <x_1,y_1>, <x_2,y_2>, \ldots\ldots\ldots <x_N, y_M>,$$

Such that x_i is the feature vector which belongs to i^{th} and y_i is its output predicted value.

E. *Unsupervised Technique*

Unsupervised learning is a machine learning technique where algorithms analyze unlabeled data to discover patterns, structures, or relationships without explicit guidance. It is used for tasks such as clustering, dimensionality reduction, and anomaly detection. By exploring the inherent structure within the data, unsupervised techniques uncover valuable insights. Clustering group's similar data points together, dimensionality reduction reduces the number of variables while preserving relevant information, and anomaly detection identifies rare or abnormal instances. Unsupervised learning plays a crucial role in exploratory data analysis and provides a foundation for further analysis and decision-making in various domains.

III. NETWORK TRAFFIC CLASSIFICATION MODEL

In this section, we will discuss the model for network traffic classification, which involves a step-by-step process outlined in Figure 3. This method guides you on how to utilize machine learning techniques for identifying and categorizing unknown network traffic classes through network traffic classification techniques.

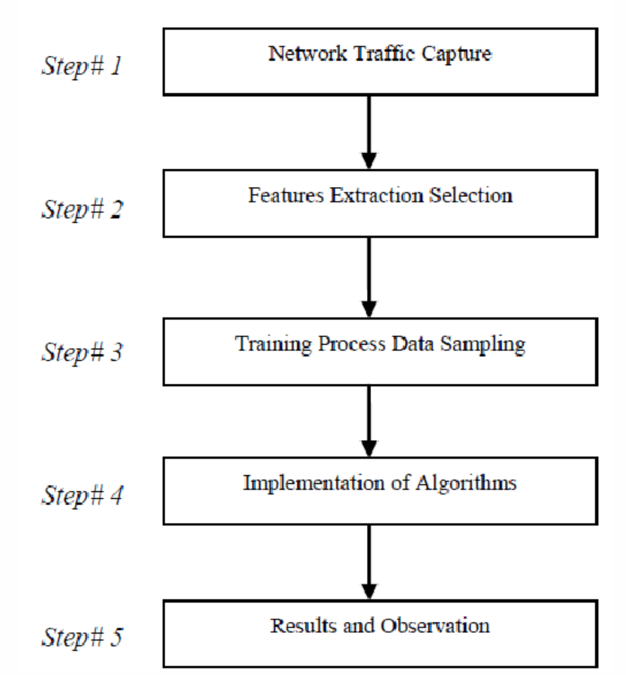

Step# 1
Network Traffic Capture
Step# 2
Features Extraction Selection
Step# 3
Training Process Data Sampling
Step# 4
Implementation of Algorithms
Step# 5
Results and Observation

Figure 3. Network traffic classification model.

A. *Network Traffic Capture*

The initial and essential stage in network traffic classification is data collection, which entails capturing real-time network traffic using different tools. Tcpdump is commonly used for this purpose, while Wireshark is employed for capturing and analyzing packets. During this step, the traffic of specific applications like WWW, DNS, FTP, P2P, and Telnet is recorded for a minute. The captured data is subsequently utilized in subsequent stages of network traffic classification, including model construction and training of machine learning algorithms.

B. *Feature Extraction Selection*

After collecting network traffic data, the next step in network traffic classification is feature selection and extraction. In this step, features are extracted from the captured data, such as packet duration, packet length, inter-arrival packet time protocol, etc. These features are then used to train machine learning classifiers. Feature extraction can be performed using a Perl script to extract features from the captured data set. However, the Netmate tool is commonly used for this purpose, which extracts 23 features. The dataset is saved in MS Excel and exported in CSV file format for use in Weka tool for further analysis and classification.

C. *Training Process Sampling*

In this stage of network traffic classification, the data sets are sampled for the supervised learning technique. This involves labeling the data to classify unknown network applications. The labeled data sets are then used to train the machine learning algorithms for network traffic classification. Figures 6 and 7 show the recall and precision results of the captured WWW, DNS, FTP, P2P, and Telnet applications. These figures help to evaluate the performance of the machine learning algorithms used in the classification process. Based on these figures, we can determine which applications have good recall and precision results and which ones do not. For example, DNS and WWW applications have poor recall and precision results compared to other applications.

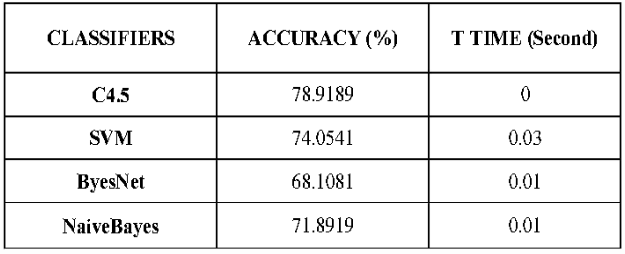
D. *Implementation of Machine Learning Algorithms*

In this step of network traffic classification, machine learning algorithms or classifiers are applied to the instances. These algorithms include supervised, unsupervised, and semi-supervised learning, which can be implemented using various tools available on the internet. Popular tools used for implementing these algorithms are MatLab and Weka classification simulation tools. The paper being referred to in this context uses Weka and applies four machine learning algorithms, namely C4.S, Support Vector Machine, BayesNet, and NaiveBayes, to build a classification model using a 10 Folder Cross-validation approach.

E. *Result and Observation*

Upon implementing machine learning algorithms, the simulation tool provides comprehensive results regarding the applied algorithms, including detailed accuracy information, training time, and recall, among others. In this study, we employed four classifiers: C4.S, Support Vector Machine, BayesNet, and NaiveBayes. Notably, the C4.S algorithm demonstrated significantly higher accuracy compared to the other algorithms. Table 3 displays the accuracy and training time results, while Figure 4 illustrates the comparison of accuracy outcomes achieved by the four machine learning algorithms.

TABLE III. ACCURACY RESULT AND COMPARISON OF MLA.



CLASSIFIERS	ACCURACY (%)	T TIME (Second)
C4.5	78.9189	0
SVM	74.0541	0.03
ByesNet	68.1081	0.01
NaiveBayes	71.8919	0.01

Table III demonstrates that the C4.S machine learning classifier outperforms other applied classifiers in terms of accuracy. Additionally, Figure 4 visually represents the accuracy results, highlighting the classifier that yields the most precise accuracy. Furthermore, Figure S displays a comparison of recall and precision results, indicating that the C4.S classifier performs exceptionally well compared to other machine learning classifiers. Its accuracy and effectiveness make it a superior choice for the given task.

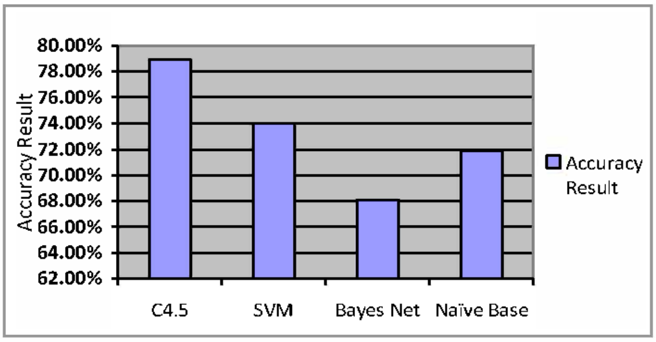


Figure 4. Accuracy result and comparison.

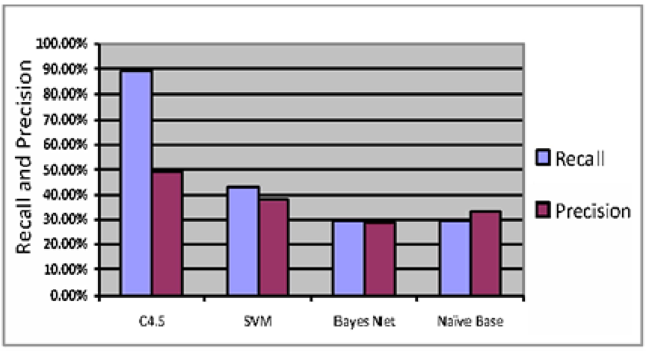


FIGURE 5. RECALL AND PRECISION OF FOUR MACHINE LEARNING

Figures 6 and 7 present the recall and precision results specifically for the captured WWW, DNS, FTP, P3P, and TELNET applications. These figures provide insights into the performance of each application in terms of recall and precision. It is evident that DNS and WWW applications exhibit notably poor recall and precision results compared to the other applications. This indicates that the performance of DNS and WWW applications may be subpar in terms of accurately identifying relevant instances and minimizing false positives.

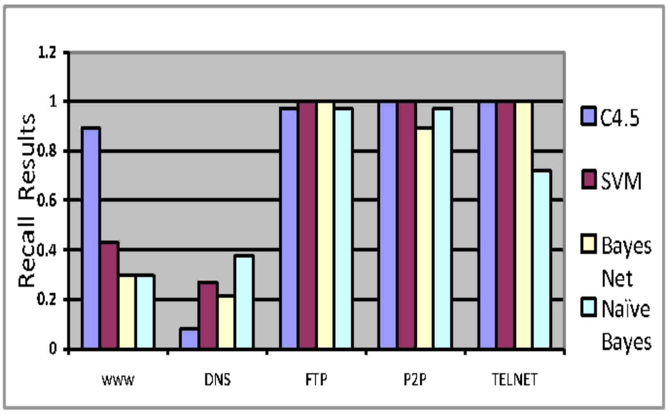


Figure 6. Recall of four machine learning classifiers on five applications.

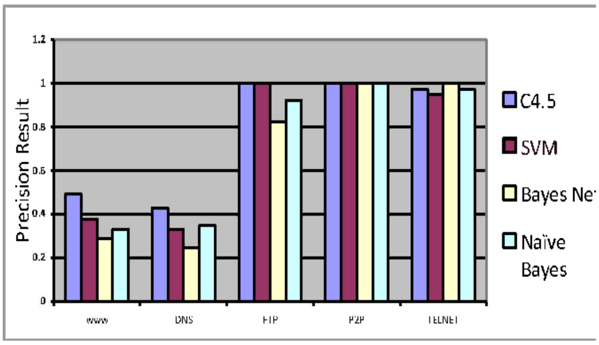


Figure 7. Precision of four machine learning classifiers of five applications.

IV. CONCLUSION

This paper aims to provide guidance to new researchers or network operators on applying network traffic classification techniques using machine learning algorithms. We discuss the process of classifying unknown applications and effectively managing network performance. Additionally, a comparative analysis of four machine learning classifiers is performed. The paper begins by presenting Network Traffic Classification Techniques, including Port-Based, Payload-Based, and Machine Learning-Based techniques, along with their limitations. We then outline a structured model for network traffic classification, starting from traffic capture to the final results. To conduct the comparative analysis, we capture one minute of traffic data for five applications: WWW, DNS, FTP, P3P, and TELNET, using the Wire Shark tool. We extract 23 features from the captured data using the Netmate tool. Subsequently, the traffic is classified using four machine learning algorithms. The experimental results demonstrate that the C4.5 decision algorithm achieves the highest accuracy compared to the Support Vector Machine, BayesNet, and NaiveBayes machine learning classifiers. These findings provide valuable insights for effective network traffic classification and management.

REFERENCES

[1]. Thuy Introduction to Network Traffic Classification. (n.d.). http://www.cisco.com/c/eu/us/td/docs/nsite/ .. ./chap05.pdf.

[2]. Internet Assigned Numbers http://www.iana.orglassignments/port2008. Authority (lANA), numbers, as of August 12,

[3]. T. Nguyen, and G. Armitage, A Survey of Techniques for Internet Traffic Classification using Machine Learning, IEEE Surveys and Tutorials, 10(4), pp. 56-76, 2008.

[4]. Pawel Foremski, On different ways to classify Internet traffic: a short review of selected publications Theoretical and Applied Informatics, 2013.

[5]. T. Karagiannies, A. Broido, and M. Faloutsos, 'Transport layer identification of P2P traffic," Proc. of ACM SIGCOMM IMC, August, 2004.

[6]. T. Karagiannies, A. Broido, N. Brownlee, K. Claffy and M. Faloutsos, "File-sharing in the internet: a characterization of p2p traffic in the backbone," Proc. of ACM SIGCOMM IMC, August, 2004.

[7]. A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in Porco Of PAM Conf., March, 2005.

[8]. 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.

[9]. Arthur Callado, Carlos Kamienski, Geza Szabo, Balazs Peter GerYo, Judith Kelner,Stenio Fernandes ,and Djamel Sadok. "A Survey on Internet Traffic Identification," IEEE Communications Survey & tutorials, Vol. II, No. 3, pp. 37-52, Third Quarter 2009.

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

[11]. T. Auld, A. Moore, and S. Gull, \Bayesian neural networks for Internet tracclassication," IEEE Transactions on Neural Networks, vol. 18, no. 1,2007.

[12]. http://www.mathworks.com/help/stats/supervised-learning-machinelearning-workflow-and-algorithms.html

[13]. Waikato Environment for Knowledge Analysis (WEKA) 3.4.4, http://www.cs.waikato.ac.nz/ml/weka!.

[14]. Knowledge Analysis http://www.mathworks.com/downloads/ Matlab.

[15]. To capture online traffic, Wire shark tool, Application: http://www.wireshark.org

[16]. Kuldeep Sing, Sunil Agrawal, "Comparative Analyssis of Five Machine Learning Algorithms for IP Traffic Classification" IEEE International Conference on 2011 sEmerging Trends in Network and Computer Communication.

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

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

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

[20]. Introduction to NetMate toll, download infonnation https://dan.arndt.calnims/calculating-flow-statistics-usingnetmatel comment -page-II

[21]. Cao, ,lie, et al. "Network Traffic Classification Using Feature Selection and Parameter Optimization." Journal of Communications 10.10 (2015).