1. ABSTRACT

The main aim of Network Traffic Classification is to classify the network traffic coming from different applications by analyzing the data packets that were received. Network Traffic is nothing but the data traffic i.e., the amount of data flowing in a particular network. Nowadays there's a widespread use of encryption techniques in network applications and network traffic classification has become a major challenge for managing the network. Network Traffic Classification is now a very important task for Internet Service Providers in order to know the type of applications flowing in a network and is used to analyze the different types of applications in a network. Nowadays the most common technique we have used is Machine Learning Based Techniques because it has given more accurate and effective results. We used four different machine learning algorithms on encrypted data and finally we got 99.8% Accuracy for Decision Tree Algorithm, 92% for Random Forest Algorithm,99% for Naive Bayes Algorithm and 87% for KNN Algorithm respectively.Network Traffic Classification is a central topic nowadays in the field of computer science. It is a very essential task for internet service providers (ISPs) to know which types of network applications flow in a network. Network Traffic Classification is the first step to analyze and identify different types of applications flowing in a network. There are many traditional techniques to classify internet traffic like Port Based, Payload Based and Machine Learning Based techniques. The most common technique used these days is Machine Learning (ML) technique, which is used by many researchers and has very effective results. In this project we attempt to implement a machine learning approach to classify Encrypted network traffic. This project, aims at building models like Decision Tree, Random forest ,knn for Encrypted network traffic classification. Nowadays there is a widespread use of encryption techniques in network applications and network traffic classification has become a major challenge for managing the network. In our project we'll use Machine Learning Based Techniques to analyze the traffic using algorithms like Decision Tree, Random Forest, KNN, Naive Bayes.

Table of Contents: Page-No

LIST OF FIGURES

2. INTRODUCTION

Data encryption has become the primary means of maintaining the privacy of Internet communications. According to data released in April 2020 by Statistica, 63 percent of organizations use the Transport Layer Security (TLS) and Secure Socket Layer (SSL) cryptographic protocols extensively. Another 23 percent use them partially.

The introduction of network traffic encryption has significantly improved communication security and user privacy. When using technologies, like Transport Layer Security (TLS), most internet users assume that third parties cannot gain access to their communications and companies rest assured that their transactions are safe from interference and eavesdropping.

Encryption plays an essential role in data security and privacy, but it also provides cybercriminals with an efficient mechanism for distributing malware. That's why you need security tools that are capable of inspecting encrypted network traffic.

To that purpose, research and methods are evaluated through the following essential use cases:

- Application identification;
- Network analytics;
- User information identification;
- Detection of encrypted malware;
- File/Device/Website/Location fingerprinting;
- DNS tunnelling detection.

Traffic classification is an automated process which categorizes computer network traffic according to various parameters (for example, based on port number or protocol) into a number of traffic classes. Traffic classification methods using flow and packet based measurements have been previously researched using various techniques ranging from automated machine learning (ML) algorithms to deep packet inspection (DPI) for accurate application identification. Nowadays there's more demand for encryption techniques in network applications, and encrypted network traffic has become a huge challenge for network management. Studies on encrypted traffic classification not only help to strengthen the network service quality, but also assist in enhancing network security. Here we have introduced the essential information of encrypted traffic classification, emphasizing the influences of encryption on present classification methodology. Then, we have described all the challenges and recent advances in the encrypted traffic classification research. This has presented a challenge for traffic measurement, especially for analysis and anomaly detection methods, which are hooked into the sort of network traffic. Next, we have looked over existing approaches for classification and analysis of encrypted traffic. First, we have described the foremost widespread encryption protocols used throughout the web. Also, We've shown that the initiation of an encrypted connection and therefore the protocol structure divulge much information for encrypted traffic classification and analysis. The purpose of knowledge encryption is to guard digital data confidentiality because it is stored on computer systems and transmitted using the web or other computer networks. In computing, encryption is the conversion of knowledge from a readable format into an encoded format which may only be read or processed after it has been decrypted. Firms of all sizes typically use encryption to guard sensitive data on their servers and databases. Encryption also provides cyber criminals with an efficient mechanism for malware distribution. Encryption provides Security for data at all times.

New work is emerging on the use of statistical traffic characteristics to assist in the identification and classification process. This survey paper looks at emerging research into the application of Machine Learning (ML) techniques to IP traffic classification - an interdisciplinary blend of IP networking and data mining techniques. We provide context and motivation for the application of ML techniques to IP traffic classification, and review 18 significant works that cover the dominant period from 2004 to early 2007. These works are categorized and reviewed according to their choice of ML strategies and primary contributions to the literature. We also discuss a number of key requirements for the employment of ML-based traffic classifiers in operational IP networks, and qualitatively critique the extent to which the reviewed works meet these requirements. The dynamic classification and identification of network applications responsible for network traffic flows offers substantial benefits to a number of key areas in IP network engineering, management and surveillance. Currently such classifications rely on selected packet header fields (e.g. port numbers) or application layer protocol decoding. These methods have a number of shortfalls e.g. many applications can use unpredictable port numbers and protocol decoding requires a high amount of computing resources or is simply infeasible in case protocols are unknown or encrypted. We propose a novel method for traffic classification and application identification using an unsupervised machine learning technique. Flows are automatically classified based on statistical flow characteristics. We evaluate the efficiency of our approach using data from several traffic traces collected at different locations of the Internet. We use feature selection to find an optimal feature set and determine the influence of different features.Timely and accurate traffic classification and application characterization are becoming increasingly important with many applications in wired and wireless networks, e.g., traffic engineering, security monitoring, and quality of service (QoS). In particular, Software Defined Networking (SDN) is a new networking paradigm that has great impact on future IP networks and 5G wireless networks. In SDN networks, application awareness is essential for functionalities such as virtual network resource slicing and fast routing. Compared to traditional classification methods such as port-based and payload-based algorithms, machine learning (ML) approaches offer a better choice in Internet traffic characterization by using payload-independent traffic statistics. In this paper, two ML algorithms, namely supervised Support Vector Machine (SVM) and unsupervised K-means clustering, are studied for traffic classification. It has been found that an overall accuracy of over 95% can be achieved. Meanwhile, the system performance can be further improved with model tuning and feature selection.

The task of network management and monitoring relies on an accurate characterization of network traffic generated by different applications and network protocols. We employ three supervised machine learning (ML) algorithms, Bayesian Networks, Decision Trees and Multilayer Perceptrons for the *flow-based* classification of six different types of Internet traffic including peer-to-peer (P2P) and content delivery (Akamai) traffic. The dependency of the traffic classification performance on the amount and composition of training data is investigated followed by experiments that show that ML algorithms such as Bayesian Networks and Decision Trees are suitable for Internet traffic flow classification at a high speed, and prove to be robust with respect to applications that dynamically change their source ports. Finally, the importance of correctly classified training instances is highlighted by an experiment that is conducted with wrongly labeled training data.

Internet traffic classification is one of the popular research interest area because of its benefits for many applications like intrusion detection system, congestion avoidance, traffic prediction etc. Internet traffic is classified on the basis of statistical features because port and payload based techniques have their limitations. For statistics based techniques machine learning is used. The statistical feature set is large. Hence, it is a challenge to reduce the large feature set to an optimal feature set. This will reduce the time complexity of the machine learning algorithm. This paper tries to obtain an optimal feature set by using a hybrid approach -An unsupervised clustering algorithm (K-Means) with a supervised feature selection algorithm (Best Feature Selection).

3.LITERATURE SURVEY

"What other people think” has always been an important piece of information for most of us during the decision-making process. The Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics — that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet. The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that is driving force for this area of interest. And there are many challenges involved in this process which need to be walked all over in order to attain proper outcomes out of them. In this survey we analysed basic methodology that usually happens in this process and measures that are to be taken to overcome the challenges being faced.

EXISTING METHODS:

3.1 Evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks.

In [2] Practical evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks. Encrypted traffic classification plays or sort of Internet traffic. Nowadays, traffic classification has become a challenging task due to an increase in the latest technologies, like traffic encryption and encapsulation, which can decrease the performance of classical traffic classification strategies. Machine learning will gain interest as a replacement direction during this field, showing the signs of future success, like knowledge extraction from encrypted traffic, and more accurate Quality of Service management. Machine Learning has now become a key tool to make traffic classification solutions in real network traffic scenarios; during this sense, the aim of this investigation is to explore the elements that allow this system to figure within the traffic classification field. Therefore, a scientific review is introduced that supports the steps to realize traffic classification by using ML techniques.

3.4 Network traffic classification using k-means clustering

Network traffic classification and application [5] identification has now become very important for IP network engineering, management and control and other key domains. Current popular methods, like port-based and payload-based, have shown some disadvantages, and therefore the machine learning based method may be a potential one. The traffic is assessed consistent with the payload-independent statistical characters. This paper introduces the varied levels in network traffic-analysis and thus the relevant knowledge in the machine learning domain, analyzing the problems of port-based and payload-based methods in traffic classification. Considering the priority of the machine learning-based method, we experiment with unsupervised K-means to improve the efficiency and performance. We adopt feature selection to seek out an optimal feature set and log transformation to enhance the accuracy. The experimental results on different datasets convey that the method can obtain up to 80% overall accuracy, and, after a log transformation, the accuracy is improved to 90% or more.

3.5 Analyzing Encrypted Network Traffic of Mobile Devices

As years have passed, [6] smartphones have come to dominate several areas that improve our lives, offering us convenience, and reshaping our daily work circumstances. Also, there are many advantages like gaming, browsing, and shopping. There will be a certain amount of traffic over the internet that belongs to the applications running over mobile devices. Applications encrypt their communication in order to maintain the privacy and security of the user's data. Now, it's been found that the number of incoming and outgoing traffic in a mobile device can result in reveal a big amount of data which will be wont to trace the activities performed by the user, researchers attempt to develop techniques to classify encrypted mobile traffic at different levels of granularity, with the objectives of performing mobile user profiling, network performance optimization, etc. This paper is employed to categorize the research works on analyzing the encrypted network traffic that are associated with mobile devices. Then, we provide a thorough review of state of the art supported the proposed framework.

3.6 Encrypted traffic classification using Machine Learning for identifying different social media applications

Social media applications such as WhatsApp, Facebook, YouTube etc. are popular representatives of encrypted traffic [7] and have grabbed big attention to communication and entertainment. Therefore, the accurate identification of them within network traffic has become a big issue to explore them in detail. In this topic, Machine Learning Techniques have shown promise in this area especially for detecting and classifying the encrypted traffic data. Therefore, this work concentrates on the challenges and the ability to use Machine Learning algorithms for social media classification from traffic. This problem statement worked on four different machine learning algorithms. i.e., support vector machine, naïve bayes algorithm, C4.5 algorithm, MLP algorithm. Features are defined by source IP, source port, destination IP, destination port, and protocol.

This problem statement worked on four different machine learning algorithms. i.e., support vector machine, naïve bayes algorithm, C4.5 algorithm, MLP algorithm. Features are defined by source IP, source port, destination IP, destination port, and protocol. All classification methods are trained using a training dataset and then tested for their performance using the test dataset. The result for classification under different ML techniques. In the case of four applications (Facebook, YouTube, Skype, and WhatsApp), the C4.5 algorithm provides a classification accuracy which is better than other ML algorithms employed in identifying the mentioned traffics. In this case, the C4.5 gave about 88.29 % accuracy of the test samples. Remarkably, the C4.5 algorithm provides the best accuracy results because the relationship between our selected features and its class is simple and other methods, such as the neural networks, will create unique network architecture and overfit the training data.

3.7 Network traffic classification and comparative analysis using machine learning algorithms

Through this technique, internet service providers can handle the overall performance of a network. There are many traditional techniques to classify internet traffic like Port Based, Payload Based and Machine Learning Based techniques. The most common technique used these days in the field is Machine Learning technique [8] Which is used by researchers and got very effective and amazing accuracy results. In this paper, we study about network traffic classification techniques step by step and real time internet data set is developed using network traffic capture tool and feature extraction tool is used to extract features from the capture traffic and then four machine learning classifiers Support Vector Machine, C4.5 decision tree, Naive Bayes and Bayes Net classifiers are applied.

The features are extracted from the captured data such as packet duration, packet length; inter arrival packet time protocol etc. Then extracted features are used to train the machine learning classifier. For feature extraction, Perl script is also used to extract the feature from a captured data set. But we use the Net mate tool. We use MS Excel for saving the dataset for Weka tool as a Comma Separated Values (CSV) file format. Experimental analysis shows that the C4.5 algorithm gives very good accuracy results as compared to other Machine Learning algorithms.

3.8 Network Traffic Classification using Machine Learning techniques for Software Defined Networks

[9] Network data gathered by the SDN controller will permit data analytics methods to analyze and apply machine learning models to customize the network management. This paper has mainly concentrated on analyzing network data and implementing a network traffic classification solution using machine learning and integrating the model in software-defined networking platforms.