IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

Network Traffic Classification on Encrypted Data using Machine Learning Approach

S Ratan Kumar[1], Ch Vinay[2], A Mounika Satya Lakshmi[3] , K Jahnavi[4], S Karthik[5]

Associate Professor, Dept of CSE, ANITS, Visakhapatnam, AP, India[1]

UG Student, Dept. of CSE, ANITS, Visakhapatnam, AP, India[2]

UG Student, Dept. of CSE, ANITS, Visakhapatnam, AP, India[3]

UG Student, Dept. of CSE, ANITS, Visakhapatnam, AP, India[4]

UG Student, Dept. of CSE, ANITS, Visakhapatnam, AP, India[5]

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.

KEYWORDS: Decision Tree, Encryption, Internet Service Providers, KNN, Naive Bayes, Network Traffic, Payload Based, Port Based, Random Forest.

I.INTRODUCTION

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.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

II.LITERATURE SURVEY

In [1] evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks are analyzed. Encrypted traffic classification played a very important role in cybersecurity as network traffic encryption has become more prevalent. Now, we haveintroduced three traffic encryption mechanisms such as IPsec, SSL/TLS, and SRTP. Since we have evaluated the performances of support vector machine, random forest, naïve Bayes, and logistic regression for traffic classification, now we haveproposed the combined approach of entropy estimation and artificial neural networks. First, network traffic can be classified as encrypted network traffic or plaintext with entropy estimation. Encrypted traffic can also be classified using neural networks. We have proposed this system using traffic packet's sizes, packet's arrival time, and direction as the neural network's input. Our combined approach had been evaluated using the dataset obtained from the Canadian Institute for Cybersecurity.

In [2] Nowadays there's a widespread use of encrypted data transport, network traffic encryption has become typical these days. This will become a challenge for traffic measurement, mainly for analysis and anomaly detection methods, which are hooked into the sort of network traffic. Next, we'velooked over existing approaches for classification and analysis of encrypted traffic. First, we've described the foremost widespread encryption protocols used throughout the internet. We had shown that the initiation of an encrypted connection and thus the protocol structure divulge much information for encrypted traffic classification and analysis. Then, we'velooked over payload and feature-based classification methods for encrypted traffic and categorized them employing a longtime taxonomy. The advantage of sort of described classification methods is that the facility to acknowledge the encrypted application protocol additionally to the encryption protocol. Finally, we made a comprehensive comparison of the surveyed feature-based classification methods and presented their weaknesses and strengths.

In [3] the purpose of Traffic analysis is used to determine all relationships, patterns, anomalies, and misconfigurations, within the network traffic. In particular, traffic classification may be a subgroup of strategies during this field that aims at identifying the application's name 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 hasgained interest as a replacement direction in this field, showing signs of future success, like knowledge extraction from encrypted traffic, and more accurate Quality of Service management. Machine Learning has become a key tool to make traffic classification solutions in real network traffic scenarios; during this sense,this investigation hasexplored 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. The main aim is to know and to spot the procedures followed by the prevailing works to realize their goals. As a result, this survey paper has found a group of trends derived from the analysis performed on this domain; during this manner, the authors expect to stipulate future directions for ML-based traffic classification.

According to [4] Network traffic classification and application 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. Here, it introduced 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'veexperimented with unsupervised K-means to improve the efficiency and performance. We've adopted 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.

[5] As years have passed, smartphones have come to dominate several areas that haveimproved 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 have encrypted 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 has resulted in revealing a big amount of data which will be wont to trace the activities performed by the user, researchers have attempted 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. It is employed to categorize the research works on analyzing the encrypted network traffic that are associated with mobile devices. Then,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

we've provided a thorough review of state of the art supported the proposed framework.

According to [6], previously work was done by the encrypted traffic classification by using a neural network model with deep and parallel network-in-network (NIN) structure by using this method for classifying encrypted network traffic, then compared with standard Convolutional neural network (CNN).in this proposed system NIN Adopts a micro network after each and every convolutional layer for local modeling. Besides, NIN utilizes a GAP Layer(global average pooling) instead of full connected convolution layer before final classification. By using GAP layer it has reduced the overfitting and model parameterssignificantly.The dataset "ISCX VPN-non-VPN" used here was captured by the author in their daily life. Besides, the dataset consists of a pcap file corresponding to various network applications. By using this dataset they've trained the CNN and NIN by comparing these two algorithms. NIN has given the best accuracy when compared to CNN. The results obtained by the Algorithms of f1.score is 0.983 and 0.985 achieved for traffic characterization and application identification respectively. Here, the future works left with the parallel decision strategy can further improved the accuracy of using a single NIN model for encrypted network traffic classification.

[7] Social media applications such as WhatsApp, Facebook, YouTube etc. are popular representatives of encrypted traffic 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 has concentrated on the challenges and the ability to use Machine Learning algorithms for social media classification from traffic traces. 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 has provided a classification accuracy which is better than other ML algorithms employed in identifying the mentioned traffic. In this case, the C4.5 gave about 88.29 % accuracy of the test samples. Remarkably, the C4.5 algorithm has provided the best accuracy results because the relationship between our selected features and its class is simple and other methods, suchas the neural networks, havecreated unique network architecture and overfit the training data.

III.PROPOSED SYSTEM

We'vefound the best suitable algorithm from the set of machine learning algorithms which provide best accuracy for the traffic flows taken in the dataset. Here, we have taken IP Network Traffic flows labeled with 75 Apps(Dataset).This dataset contains 87 features(like source ip, destination-ip, source port, protocol, protocol name).Each instance hold the information of an ip flow(Traffic flow) generated by network device .But mainly this dataset have gone a step further by generating machine learning models capable of detecting specific applications such as windows-update, amazon, https etc. After the data preprocessing stage, the dataset has applied to the set of algorithms(Decision Tree, Random Forest, Naïve Bayes, KNN).

After we've executed the code for each algorithm, we'vegot different accuracy rates for each of those algorithms then we'veselected the algorithm which has a higher accuracy percentage for the traffic flow taken in our dataset. Traffic flow is nothing but the flow of packets from source to destination i.e., flow of information from host computer to destination computer. The more we got the accuracy percentage the more accurate the algorithm is. So, we've predictedan algorithm which shows a higher accuracy percentage for the dataset we've chosen. Network traffic classification has played a very important role in categorizing the network traffic based on various parameters which was divided into a number of traffic classes and also used to prioritize network traffic for various purposes. Network traffic also known as data traffic which showed the information flow between source and destination.

IV.ARCHITECTURE

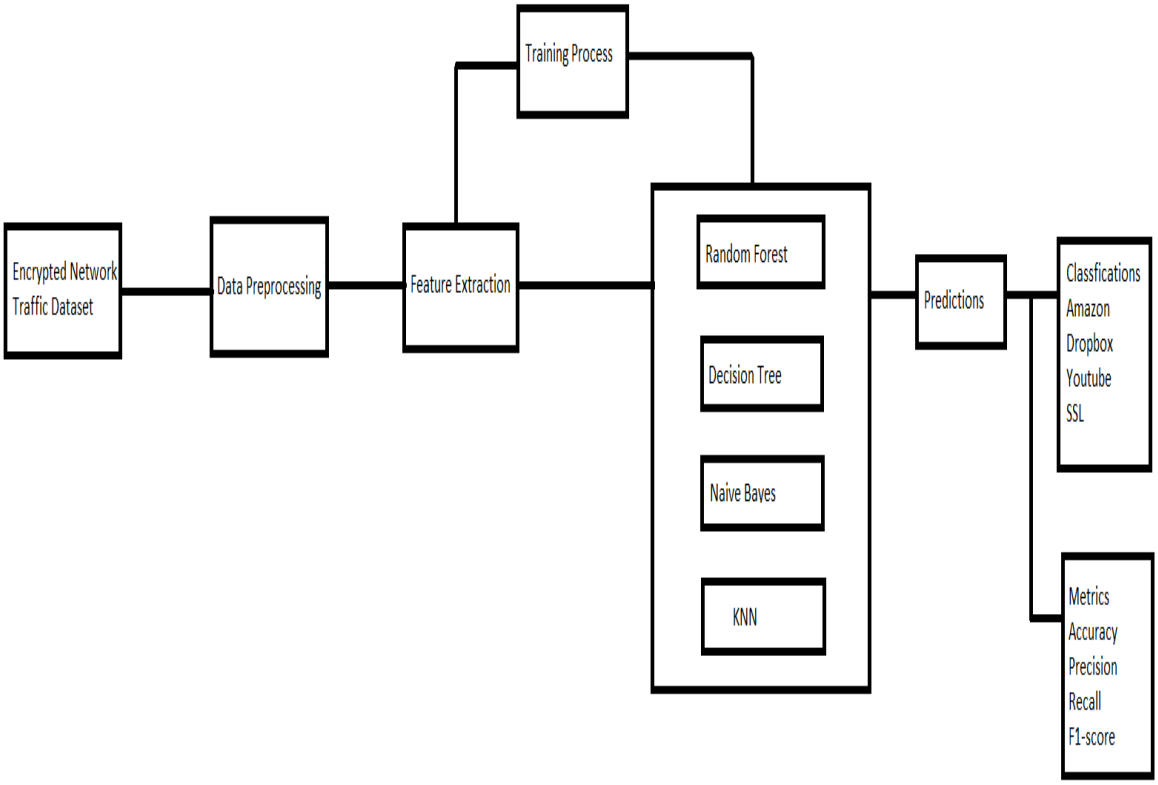
The first step is we've taken the encrypted network traffic dataset and next we have extracted the features from the dataset and next we have divided the dataset into two parts i.e.,70% training data and 30% testing data and we had performed classification on the dataset by taking 11 classes like Amazon, Dropbox, Youtube, Facebook etc., using four different machine learning algorithms: Decision Tree, Random Forest, Naive Bayes and KNN . Then, predictions are calculated using testing data and finally we've obtained the accuracy for each Algorithm.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

Fig.1. Architecture



V.RESULTS

CLASSIFIER	ACCURACY
Random Forest	92%
Decision Tree	99.8%
Naïve Bayes	99%
K-Nearest Neighbor	87%

Table 1.1 Accuracy Results

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

Random Forest:

Classes	precision	Recall	F1 .Score
AMAZON	0.97	0.97	0.97
CONTENT_FLASH	1.00	0.99	0.99
DROPBOX	0.99	0.86	0.92
FACEBOOK	0.97	0.91	0.94
GOOGLE	0.89	0.93	0.91
HHTP_CONNECT	0.84	0.99	0.91
MICROSOFT	1.00	0.99	0.99
SSL	0.92	0.99	0.95
WINDOWS_UPDATE	0.99	0.94	0.97
HTTP	0.98	0.98	0.98
YOUTUBE	0.96	0.84	0.90

Table 1.2 Random Forest

Decision Tree:

Classes	precision	Recall	F1 .Score
AMAZON	1.00	1.00	1.00
CONTENT_FLASH	1.00	1.00	1.00
DROPBOX	1.00	1.00	1.00
FACEBOOK	1.00	1.00	1.00
GOOGLE	1.00	1.00	1.00
HTTP	1.00	1.00	1.00
HHTP_CONNECT	1.00	1.00	0.98
MICROSOFT	1.00	1.00	1.00
SSL	1.00	1.00	1.00
WINDOWS_UPDATE	1.00	0.67	1.00
YOUTUBE	1.00	1.00	1.00

Table 1.3 Decision Tree

Naïve Bayes:

Classes	precision	Recall	F1 .Score
AMAZON	1.00	1.00	1.00
CONTENT_FLASH	1.00	1.00	1.00
DROPBOX	1.00	1.00	1.00
FACEBOOK	1.00	1.00	1.00
GOOGLE	1.00	1.00	1.00
HTTP	1.00	1.00	1.00
HHTP_CONNECT	1.00	1.00	1.00
MICROSOFT	1.00	1.00	1.00

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906037 |

SSL	1.00	1.00	1.00
WINDOWS_UPDATE	1.00	1.00	1.00
YOUTUBE	1.00	1.00	1.00

Table 1.4 Naive Bayes

Knn:

Classes	precision	Recall	F1 .Score
AMAZON	0.89	0.94	0.92
CONTENT_FLASH	0.97	0.99	0.98
DROPBOX	0.87	0.89	0.88
FACEBOOK	0.91	0.92	0.91
GOOGLE	0.70	0.73	0.72
HHTP_CONNECT	0.83	0.84	0.84
MICROSOFT	0.96	0.93	0.95
SSL	0.92	0.90	0.95
WINDOWS_UPDATE	0.94	0.93	0.94
HTTP	0.99	0.97	0.98
YOUTUBE	0.76	0.73	0.75

Table 1.5 Knn

VI. CONCLUSION

Thus, we can say that encrypted network classification will play an important role in categorizing the network traffic in many ways. Each sample in the dataset determines the packet flow thus we can determine the packet flow from host computer to destination computer. Finally, we have obtained the following results i.e., 100% accuracy for Decision Tree Algorithm, 92% for Random Forest Algorithm, 99% for Naive Bayes Algorithm and 87% for KNN Algorithm.

REFERENCES

[1] B. Anderson, S. Paul, and D. McGrew, Deciphering malware's use of TLS without decryption, arXiv, 2016.
[2] B. Anderson, S. Paul, and D. McGrew, malware's use of TLS without decryption, arXiv, 2018.
[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification using machine learning", IEEE Commun. Surveys Tuts., vol. 10, no. 4, pp. 56-76, 4th Quart. 2008.
[4] Yingqiu Liu, Wei Li, Yunchun Li Second international multi-symposiums on computer and computational sciences (MSCCS 2007), 360-365, 2007.
[5] S. Gai, K. McCloghrie, S. Mohaban, Method and apparatus for identifying network data traffic flows and for applying quality of service treatments to the flows, uS Patent 6,651,101 (Nov. 18 2003).
[6] Z. Bu, B. Zhou, P. Cheng, K. Zhang and Z. -H. Ling, "Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models", in IEEE Access, vol. 8, pp. 132950-132959.
[7] Furla AL-obaidy, Shadi Momlahen, Md.Faysal Hassian and Farah Mohammadi Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, Canada 2019 IEEE Canadian conference of electrical and computer engineering (CCECE).

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details