Detection of DNS over HTTPS Tunneling using Random Forest Supervised Learning.

MSc Research Project

Cybersecurity

Tejaswi Pednekar

Student ID: 21101094

School of Computing

National College of Ireland

Supervisor: Niall Heffernan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student:	Tejaswi Sharad Pednekar		
Student ID:	21101094…	**Year:**	2021-22
Programme	Cyber Security		
Module:	Msc Research		
Supervisor:	Mr. Niall Heffernan		
Submission Due Date:	19th September 2022		
Project Title:	Detection of DNS over HTTPS Tunneling using Random Forest Supervised Learning.		
Word Count:	4128 **Page Count** 16		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Tejaswi Pednekar

Date: 15-08-2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Title

Tejaswi Pednekar

X21101094

Abstract

As a network protocol, Domain Name System (DNS) is prone to a number of security flaws. For address some of DNS's weaknesses, a new protocol called DNS over HTTPS (DoH) is being developed to increase privacy and guard against certain persistent assaults. To avoid eavesdropping and man-in-the-middle attacks, the DoH protocol encrypts DNS requests for the DoH client and sends them over a tunnel. This study paper thoroughly investigates these security flaws, offers a taxonomy of probable DNS attacks, examines the security features of the DoH protocol, and categorises DNS attacks applicable to DoH. I simulated DoH tunnels to attain these goals.

Keyword: Machine Learning, Random Forest, DOH, Tunneling

1 Introduction

The Domain Name System (DNS) is a distributed hierarchical registry that maps domain names to the IP needed by communication protocol such as IPv4 and IPv6. The DNS protocol simplifies web accessibility, which contributes to the digital country's rapid development. [11]Like a result, DNS is a critical component of the TCP/IP stack and serves as the Web's address book. The DNS protocol, on the other hand, has security vulnerabilities , as well as the system may be targeted utilizing particular developed methods such as DNS Amplification & Reflection, Resource Utilization, and Domain Name system (dns Injection . DNS flaws pose a significant security risk to web users. A basic example is redirecting a DNS request to a bogus malicious website in order to steal user passwords and sensitive information. [10]

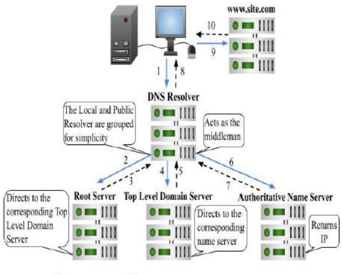

www.site.com
10
9
1
8
DNS Resolver
The Local and Public Resolver are grouped for simplicity
Acts as the middleman
2
3
4
5
6
7
Root Server
Top Level Domain Server
Authoritative Name Server
Directs to the corresponding Top Level Domain Server
Directs to the corresponding name server
Returns IP

Fig 1 DNS Procedures [14]

As both a result, the DNS protocol's security architecture must be modified to prevent unwanted access and thereby increase network privacy and security. [12] An Intrusion Detection System (IDS) is essential for monitoring the traffic of internet-connected devices and detecting DoH traffic assaults in a network architecture. Intrusion detection was characterized as "the process of monitoring and analyzing events happening in a computer system or network for indicators of intrusions, identified as efforts to undermine the privacy, authenticity, or accessibility of a system or device." [2] [13]

DNS traffic is encrypted by DoT using the Transfer Layer Security (TLS) protocol. As shown in Figure 2, a TLS connection is formed between the server and the client on port 853 [6] prior to any DNS lookups. The client requests the server and obtains a certificate to start the session. A certificate authority then validates the certificate supplied by the server. But if it all goes as planned, the client and server will share encryption keys. Once the TLS session is established, DNS requests and replies are delivered across the encrypted channel, making it difficult for a third party to access the contents of the inquiries. DoT, with exception of DNS, protects against man-in-the-middle attacks/ open public wifi.

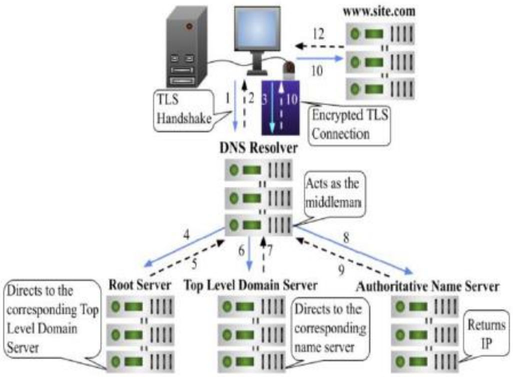


Fig 2: DOT [14]

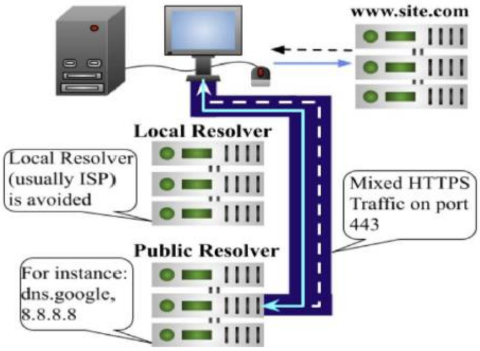

www.site.com
Local Resolver
Local Resolver (usually ISP) is avoided
Mixed HTTPS Traffic on port 443
Public Resolver
For instance: dns.google, 8.8.8.8

Fig 3 DOH [14]

Since its traffic is less identifiable, the additional privacy provided by DoH might be viewed as advantageous or harmful depending on the context. A privacy-conscious user, for example, might prefer more anonymous traffic; but, because DoH traffic is mixed in with HTTPS traffic, network administrators may find it more difficult to monitor and restrict it. Because network managers can no longer use DNS filtering or monitoring as readily as they did with DoT, the DoH protocol may reduce network protection. These strategies give useful information such as the answer IP address, the originating IP address, and the query type. DoH has experienced more uptake than DoT. [15]There are currently 17 DoH operators, including Google and Cloudflare. Above fig2 and 3 Shows how the DOT and DOH portal work with the security feature. Fig1 shows the current DNS protocol which can cause the lot of Cyber-attacks.

As we understand, what is DNS what is DOH (DNS over HTTPS) and what is DOT. For this paper, we will more focus in part of DOH as it work HTTPS, which is “*HYPERTEXT TRANSFER PROTOCAL SECURE*.” Which uses for secure the communication channel between two website using port number 443. This dissertation is divided into three phases the first stage is a presentation to the subject, the second stage is a literature review and its results are presented, and the last step is the creation, evaluation, and documentation of the entire work.

2. Related Work and Background

This part of the research will give ideas regarding related research work, and how this will helpful for technology and better ideas of improvement. This section will more focus on previous studies and investigations carried out. Doh's analysis and investigation have been shown below.

In practically all classification measures, the LGBM and XGBoost algorithms beat some other techniques. Outside of 4000 test datasets, LGBM algorithms falsely labeled one DoH traffic test as non-DoH. the source is the most important attribute for distinguishing DoH data from non-DoH traffic among the 34 derived from the CIRA-CIC-DHSBrw-2020 dataset.[2] This research concentrates on the features of encrypted traffic analysis, particularly the accurate identification of DoH. The goal is to use machine learning to determine what information (if any) can be extracted from HTTPS extended IP traffic data. They compared five prominent machine learning algorithms to determine the top DoH classifiers.[5] Due to the resemblance with that other request/response, the suggested ML system cannot identify DoH connection with a single instance. Numerous studies have been published on DNS attack techniques and defenses, as well as the study topic of DNS tunnel detection, has lately gained special attention. Although domain name resolution based on DNS is among the most basic and essential services available on the Internet, cybercriminals employ DNS properties to construct tunnels.[6] The DNS tunnel is a typical technique used by attackers to build command and control nodes and to extract relevant data off networks.[7] P. Yang et al. [38] attempted to discover DNS eavesdropping using a stacking model in recent articles on DNS tunnel detection. A variety of programs, namely dns2tcp, dnscat2, DeNiSe, and Heyoka, created the DNS traffic. They employed a stacking model, which is a combination of three techniques (K-nearest neighbors (KNN), support vector machine (SVM), and random forest).[8] By examining network traffic, DNS tunnels were also discovered. in this research, they used for DNS model whereas in this research paper DOH model. [8] One purpose of this article was to discover DNS tunneling in DNS over HTTPS networks. A real-time dataset is created and gathered by utilizing a specially deployed server and python scripts for tunneling simulation. To represent the simulated requests as tunneling or not tunneling, a dataset was compiled. The extensive feature analysis indicated the most valuable elements for achieving the desired result.[9]

Since DoH is intended to preserve the security of the IP address-to-website matching process, it is vital to categorize malicious DoH traffic. We suggested employing machine learning and deep learning approaches to detect suspicious DoH traffic in this research. They conducted comprehensive testing to evaluate the effectiveness of various learning-based strategies. To demonstrate the performance of various approaches, i used four metrics: precision, recall, accuracy, and F1-score. I discovered that model setup and tweaking were crucial for performance. Furthermore, we observed that Random Forest and Decision Tree models outperformed traditional Machine and Deep Learning models in classification results. To the best of our knowledge, there is no published study that examines numerous packet-level information of DoH and HTTPS traffic with the purpose of distinguishing DoH from standard HTTPS with high accuracy. Various ML models and potential feature vectors are evaluated in

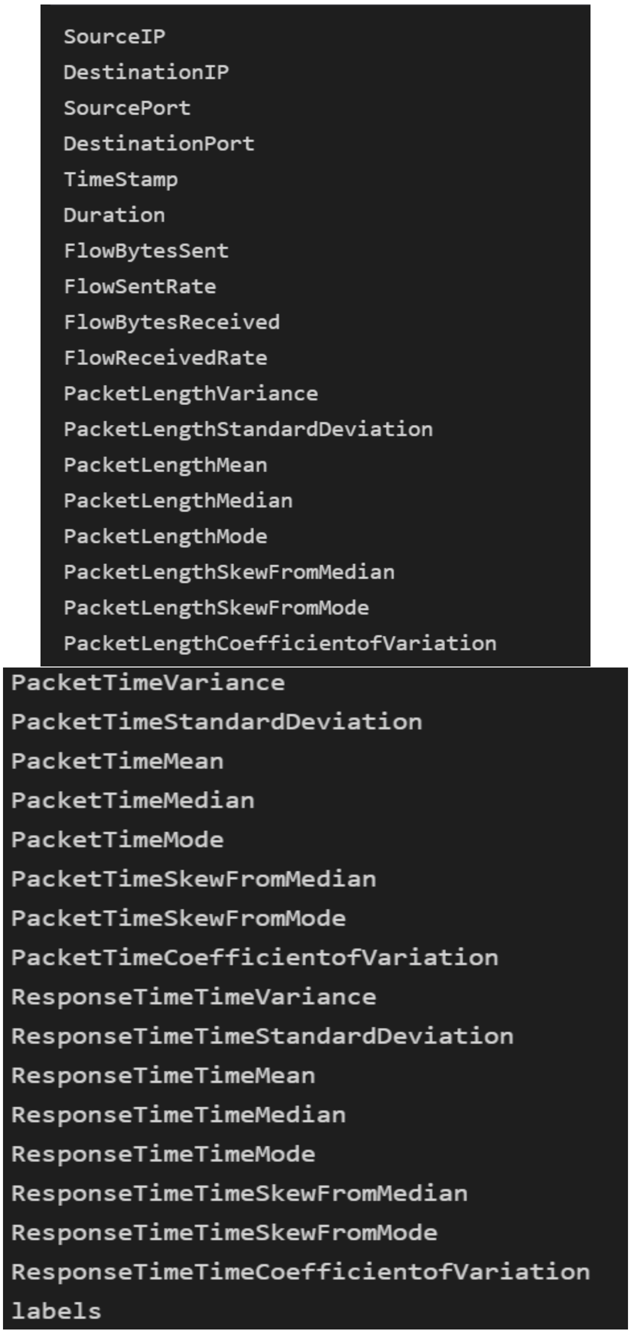
our studies. Furthermore, our research focuses on differentiating various DoH clients (applications) based on the social behaviors indicated by our analyzed feature vector.
With all the above research work my question is What are the elementary contrast in tunneling in the middle of DOH and Non-DOH concerning supervised learning. This work center on the Analysis of DNS over HTTPS, the reason for using this DOH as we used DNS everywhere to connect the world, with that securing the network is a priority in this world as many cyber-attacks are happening which I will check the high accuracy in Random Forest while comparing with other models. Taking into account all of the prior work, Random Forest and Bagging Techniques for model implementations were employed in this study and produced the predicted results, which are detailed in the following sections of this report.

3 Research Methodology

This section concentrates on the following research method for anticipating understanding in data, which incorporates machine-learning techniques on datasets received from previous studies as well as from external resources. It also underlines several essential aspects of Data reasoning and Data mining. This method is also called KDD which means knowledge discovery of Data.
In this research, I study the KDD technique has been accompanying which is most wide-ranging and acceptable for data mining.KDD is used for analysis, concentrating on massive data sets and databases, and generation of information for decision-making, Clustering depends on the discovery and visual documentation of previously unknown groups of evidence. [1] Numerous scientific endeavors, involving mathematics, cybernetics, genetics, and marketing, benefit from data mining approaches. The procedure is iterative at each level, indicating that returning to past acts may be necessary.

3.1 Data Selection:

The data set I chose for work is publicly available known as Canadian Institute for Cybersecurity (CIC) funded by the Canadian Internet Registration Authority (CIRA) (CIRA-CIC-DoHBrw-2020). [3]The researchers of this dataset recorded both good and bad DoH traffic as well as non-DoH traffic, and they utilized a two-layered classification system to separate them. Data instances are categorized as DoH or non-DoH in the first layer, and the second layer receives the data samples classified as DoH. The classification of data samples as benign-DoH or malicious-DoH occurs at the second layer. To provide further context for this information and to explain how it was created, DoH traffic, both benign and malicious, was created by reaching the top 10,000 Alexa websites using browsers and DNS tunneling software that support the DoH protocol, respectively.[4]



```
SourceIP
DestinationIP
SourcePort
DestinationPort
TimeStamp
Duration
FlowBytesSent
FlowSentRate
FlowBytesReceived
FlowReceivedRate
PacketLengthVariance
PacketLengthStandardDeviation
PacketLengthMean
PacketLengthMedian
PacketLengthMode
PacketLengthSkewFromMedian
PacketLengthSkewFromMode
PacketLengthCoefficientofVariation
PacketTimeVariance
PacketTimeStandardDeviation
PacketTimeMean
PacketTimeMedian
PacketTimeMode
PacketTimeSkewFromMedian
PacketTimeSkewFromMode
PacketTimeCoefficientofVariation
ResponseTimeTimeVariance
ResponseTimeTimeStandardDeviation
ResponseTimeTimeMean
ResponseTimeTimeMedian
ResponseTimeTimeMode
ResponseTimeTimeSkewFromMedian
ResponseTimeTimeSkewFromMode
ResponseTimeTimeCoefficientofVariation
labels
```

Fig 4 : Dataset

3.2 Dataset preprocessing

However, upon analyzing the dataset, it was discovered that there is a strong correlation seen between the Timestamp feature and whether the DoH packet is malicious or benign. As a result, the dataset developers caught the benign packet from the first network packet date through the conclusion of January 2020 and focused on the malicious packets for the remainder of the time. The usage of ML approaches will be badly impacted by this type of bias in the data set since the algorithms employed in these techniques will only use this characteristic to produce predictions that might not be accurate for the real world. Because I need to integrate the dataset in the mathematical model of these models in order to employ ML approaches, the benign and malicious records were represented in the pre-processing stage as 0 and 1, correspondingly, while the null records were denoted as -1. There are numerous ways to enter IP address from the packet source and destination but I decided to Juniper library from labelEncounter from SKlearn Library which prevents avoiding the sequence between IP addresses in case of taking integer characterization of IP address To steer clear of contract connecting the values

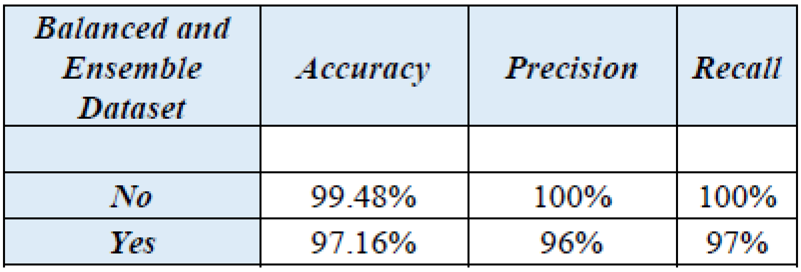
3.3 DATA TRANSFORMATION:

In this section, I used Standard Scalary library from the same library, which I used prior. After that to perform Random forest and Deep Learning techniques from Machine Learning Processes I divide the repository break into 2 parts where the majority of data was training phase and the rest to try out to apply an algorithm that shows the result and dividend show the difference between 2 machine learning techniques. Where I reduced the data for testing purposes with this I can achieve useful features to where I can denote the data that has purpose achieve tunneling with Machine learning technique. With that, I used the library which has a feature extraction model which represents special keywords and tokens with any dataset sample.

3.4 DATA MINING:

In this phase of research of had been started the procedure of picking out the verified data mining assignment such as DOH and NON-DOH data, being of Chrome and firefox data. Therefore the objective of research using the KDD process has been marked as the grading of the process. Applying the machine learning technique numerous researchers used Xbox, or gradient boosting for tunneling the DOH. To take advantage of the Random forest Model with numerous patterns has been decided. The model will give the approach with the required framework. Eventually, the procedure of Data Mining has been brought off with the dataset by searching the tunneling or traffic between DOH or nonDOH and will provide accurate accuracy.

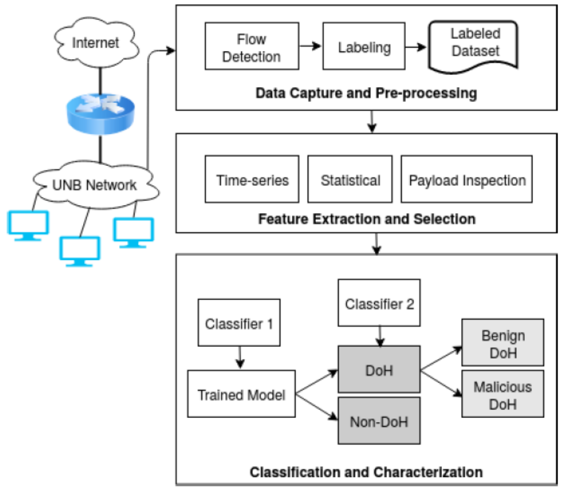
Balcnced Ensembled dataset.

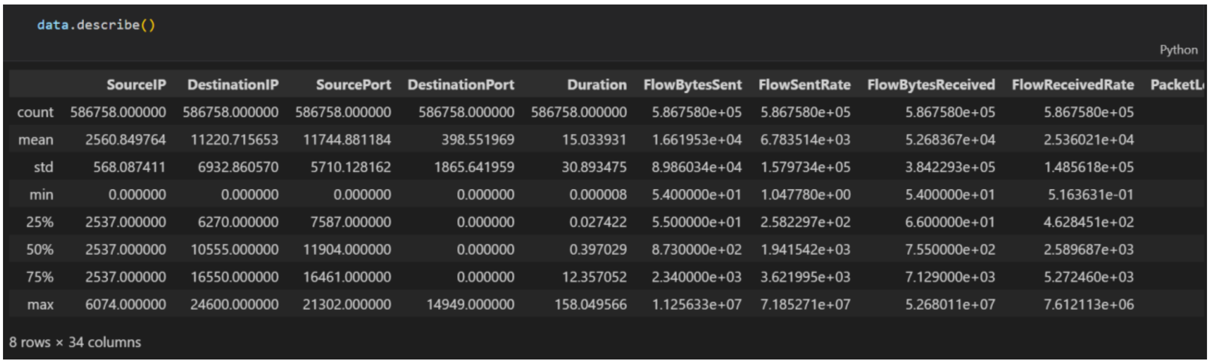


Balanced and Ensemble Dataset	*Accuracy*	*Precision*	*Recall*
No	99.48%	100%	100%
Yes	97.16%	96%	97%

4 Design Specification

Below is the DOH aricheture and And classificate how DoH technology I used to





```
data.describe()
```

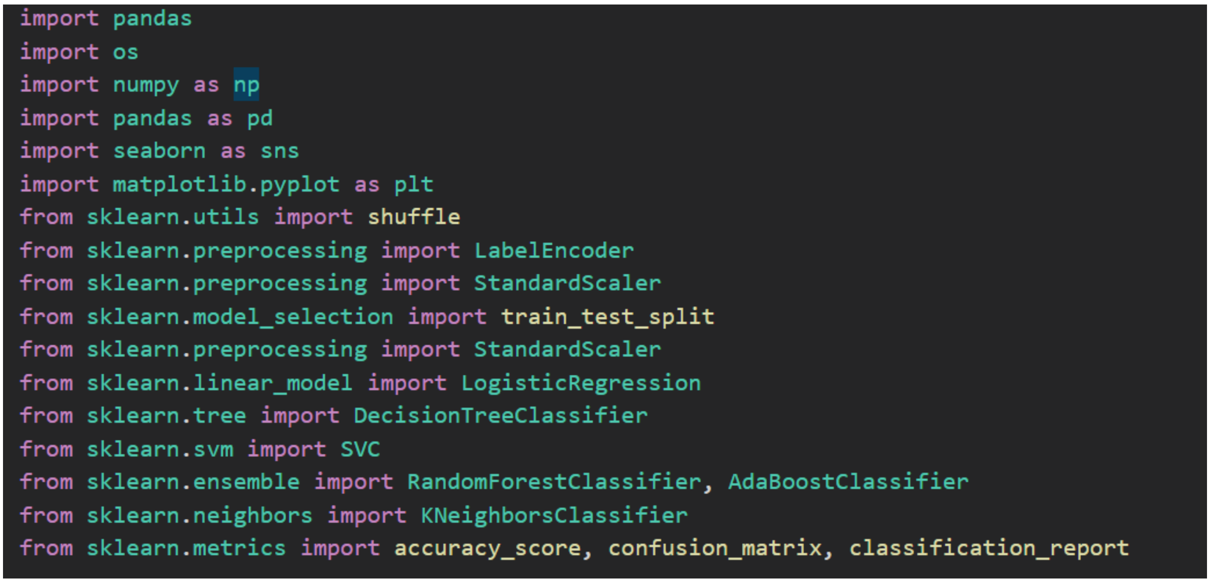
	SourceIP	DestinationIP	SourcePort	DestinationPort	Duration	FlowBytesSent	FlowSentRate	FlowBytesReceived	FlowReceivedRate	PacketL
count	586758.000000	586758.000000	586758.000000	586758.000000	586758.000000	5.867580e+05	5.867580e+05	5.867580e+05	5.867580e+05	
mean	2560.849764	11220.715653	11744.881184	398.551969	15.033931	1.661953e+04	6.783514e+03	5.268367e+04	2.536021e+04	
std	568.087411	6932.860570	5710.128162	1865.641959	30.893475	8.986034e+04	1.579734e+05	3.842293e+05	1.485618e+05	
min	0.000000	0.000000	0.000000	0.000000	0.000008	5.400000e+01	1.047780e+00	5.400000e+01	5.163631e-01	
25%	2537.000000	6270.000000	7587.000000	0.000000	0.027422	5.500000e+01	2.582297e+02	6.600000e+01	4.628451e+02	
50%	2537.000000	10555.000000	11904.000000	0.000000	0.397029	8.730000e+02	1.941542e+03	7.550000e+02	2.589687e+03	
75%	2537.000000	16550.000000	16461.000000	0.000000	12.357052	2.340000e+03	3.621995e+03	7.129000e+03	5.272460e+03	
max	6074.000000	24600.000000	21302.000000	14949.000000	158.049566	1.125633e+07	7.185271e+07	5.268011e+07	7.612113e+06	

8 rows × 34 columns

Fig5 : Data Descrition

5 Implementation

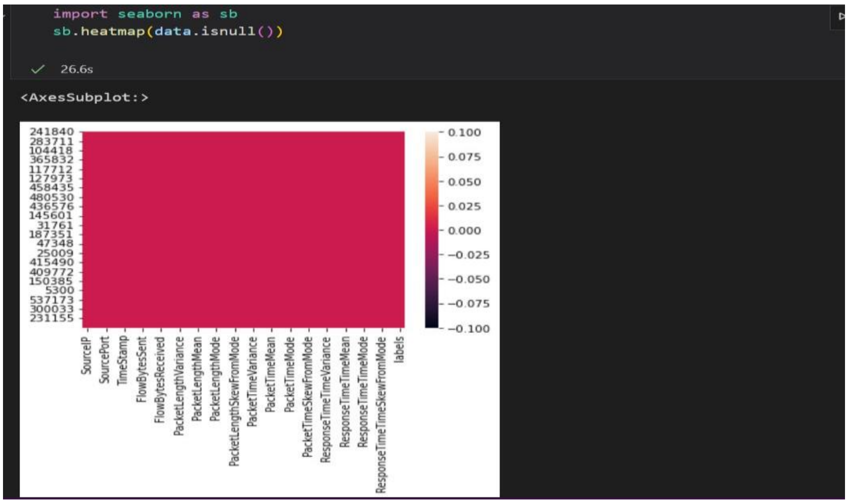
In this section of research first I import the dataset using the Jupiter pandas library with this I get to know about columns and rows. after that, I analyze the target data and get to know balanced data and unbalanced data are available in the dataset. During this step, characteristics with negligible values were removed. These comprised the Source IP, Destination IP, Packet Time Mode, and Timestamp characteristics. Some may wonder why these qualities were chosen. Source IP and Destination IP were removed from the dataset since, in practice, many programs create IP addresses at random, rendering them unsuitable for training an anomaly detection system. So rather than focusing on those traits, the machine learning model was trained using data that simulates DNS tunneling performance on the DoH protocol. A timestamp, on the other hand, was deleted since it had a high connection with the target variable. To avoid data leaking, the timestamp functionality was deleted. After that using dropna method I removed that contain duplicate values and the data which is referred to as "NA" with this I get more sorted data for testing. The label encoder component of sklearn was used to manage the category variable inside the dataset. The feature DoH was the dataset's single class label, with only two classes: True or False. As a result, those classes were changed to binary, with 0 representing False and 1 representing True. An additional pre-processing stage done to avoid a biased classification system was to equalize the targeted communications in the dataset.



```python
import pandas
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.utils import shuffle
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
```

Fig 6 :Sklearn Library and Python

That the dataset comprised more regular networking operations (355207) than DNS tunneling upon it DoH. (16257). This same dataset should be balanced in order to provide equal priority for the suggested binary classifier to work effectively. In this stage target class separation from the rest, data was taken using df. drop. for visualization used a counterplot to get a graph for labels and the count of data. I create the function that stores test accuracy, and a heatmap for the confusion matrix table that shows the accuracy and precision, and many more explanations have been provided in the evaluation. After that used a random forest library of classification using sklearn I calculate accuracy macro avg and weighted avg with using RF the test accuracy shows 99.9943%, with the same function I tested on Decision Tree Classifier with same SKLearn Liberty the accuracy shows in 99.9926%. With this shows Random forest shows the accuracy compared to another research paper.



```python
import seaborn as sb
sb.heatmap(data.isnull())
```

```
✓ 26.6s
<AxesSubplot:>
```

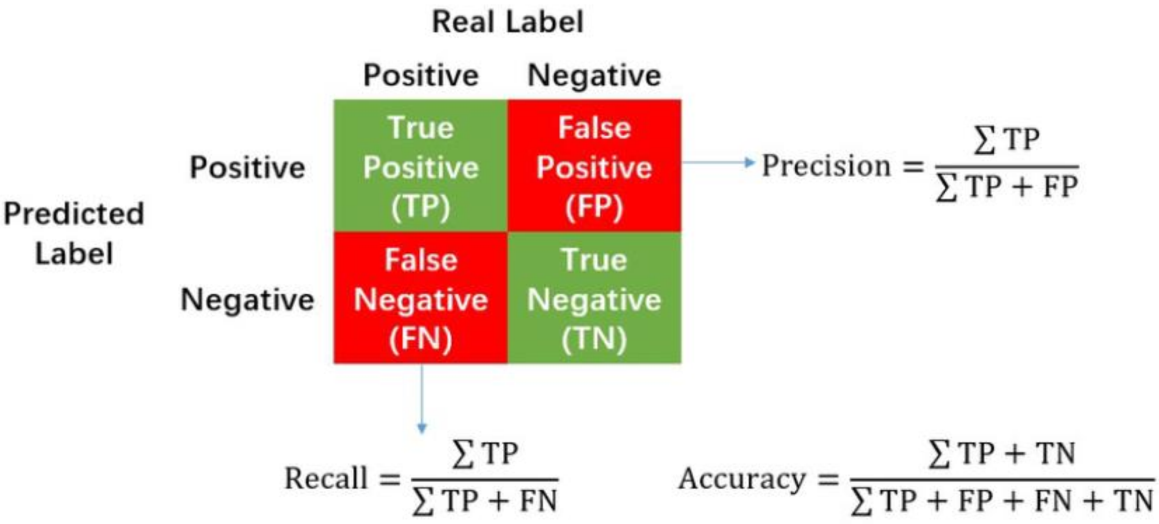
Fig7 : Graph for Null Values.

the Random Forest model to process the provided dataset using ML bagging techniques and forecast the results. The model's output is generated as a set of metrics utilizing libraries within the same sklearn package such as accuracy score, confusion matrix, and classification report.

Decision trees employ a number of sequential processes to decide whether or not to split a node into two or more sub-nodes based on the chances, prices, or even other implications of making a certain decision. A technique divides the nodes based on all available factors and then chooses the split with the most homogenous sub-nodes. Overall significance of the characteristics is extracted to ensure the algorithm's maximum accuracy. A relationship among two characteristics is examined in order to remove aspects that give the same information about the data. Because SourceIP is the most crucial characteristic in layer 1 with a feature significance of 0.67, whereas the most essential feature in layer 2 for distinguishing Benign-DoH from Malicious-DoH traffic is a newer version engineered as tan. Finally, the matplotlib library, together with the seaborn library, were utilized to perform correlation tests. This seaborn lib is used to build a graph that shows the outcomes of various models such as Decision tree Bagging and gradient boosting models with Random forest. This analysis strongly recommends the usage of the Random Forest model in this research, demonstrating that this model consistently outperforms other models in regard to accuracy.

6 Evaluation

In this section of research evaluation work has been through obtaining and determining the confusion matrix model. The use of the confusion matrix model is for deploying and visualizing the inclusive implementation of a classification model. It can compute Accuracy, Sensitivity (also known as recall), Positive predictive value, and Precision. The figure below illustrates the confusion matrix in further depth.



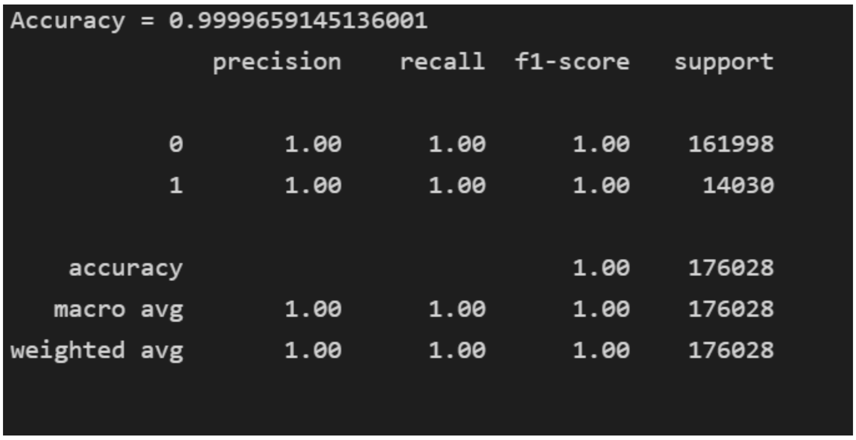
		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Fig 8 And Confusion matrices that use for the evaluation



```
Accuracy = 0.9999659145136001
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    161998
           1       1.00      1.00      1.00     14030

    accuracy                           1.00    176028
   macro avg       1.00      1.00      1.00    176028
weighted avg       1.00      1.00      1.00    176028
```

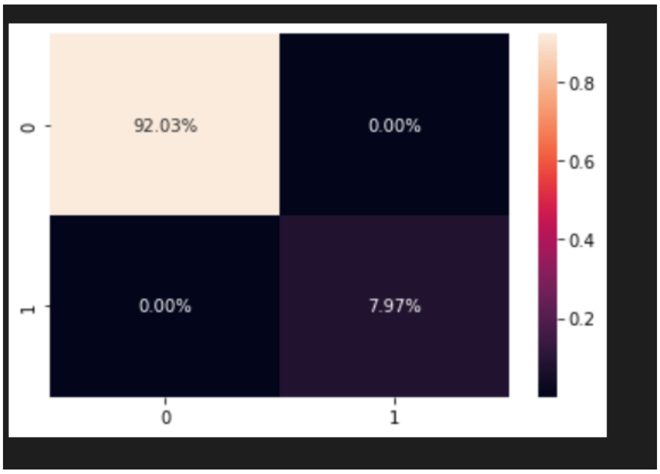
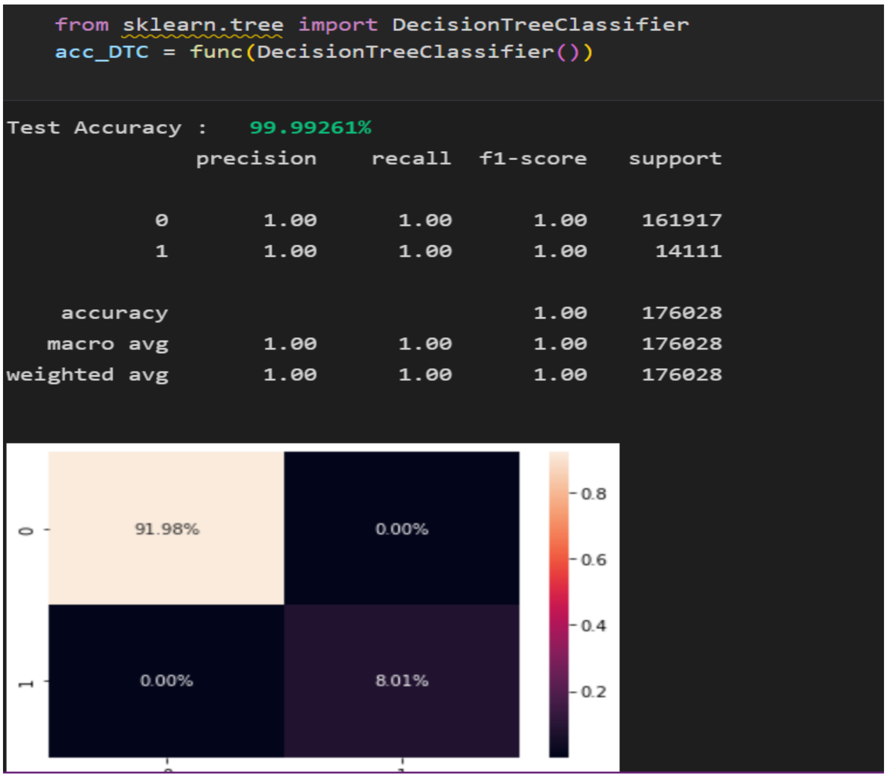


Fig 9 Random Forest Matrix



```
    from sklearn.tree import DecisionTreeClassifier
    acc_DTC = func(DecisionTreeClassifier())
```

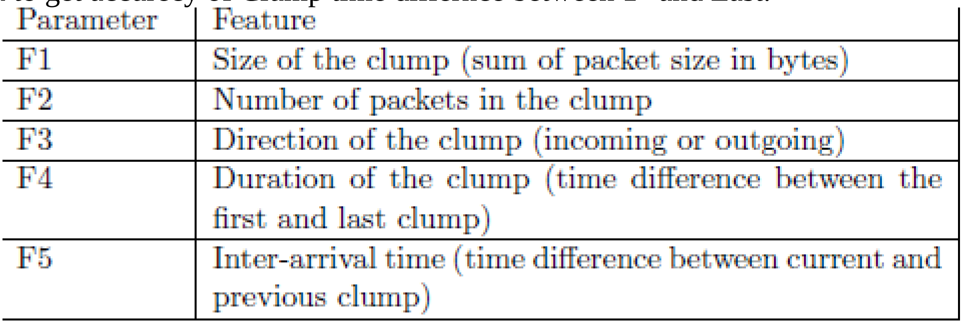
```
Test Accuracy :    99.99261%
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    161917
           1       1.00      1.00      1.00     14111

    accuracy                           1.00    176028
   macro avg       1.00      1.00      1.00    176028
weighted avg       1.00      1.00      1.00    176028
```

Fig
10: Decision Tree

The below method is used for F1 as Size of packet size
F2 Size of Number clump which gives ides how many clump or null values
F3 is used for direction of clump
F4 is used to get accurecy of Clump time differnce between 1st and Last.

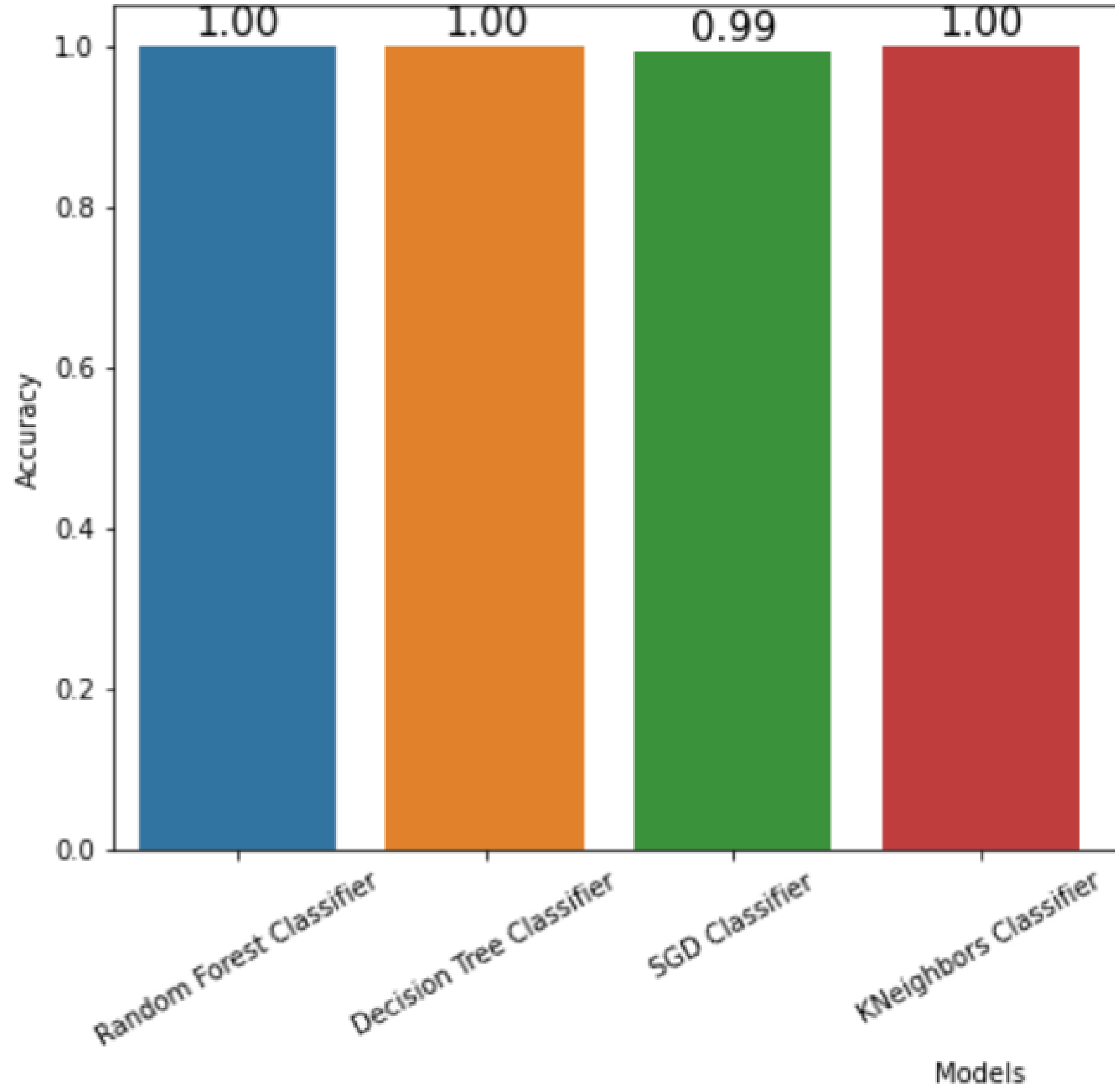


Parameter	Feature
F1	Size of the clump (sum of packet size in bytes)
F2	Number of packets in the clump
F3	Direction of the clump (incoming or outgoing)
F4	Duration of the clump (time difference between the first and last clump)
F5	Inter-arrival time (time difference between current and previous clump)

. . .

1.1 Case Study 1: Performance compared to other Algorithms

In the below tables its shows using balanced data and being and Non-DOh algorithm Machine learning profession Random Forest execute in dept in the phases of Confusion matrix like F1, recall. The Bar Chart has been indicated using seabourn lib of python which gives a clear understanding and output from a different model like Desion tree, and Random forest, SGD classifier.



Difference Between Different Machine Learning

1.2 Case Study 2:: Training time required compared to other Algorithms

Random forests are indeed an ensemble technique that combines numerous trees at the conclusion of the operation using means or the majoritarianism. A random forest classifier constructs a collection of decision trees from a randomly selected subset of the dataset and then combines the results from multiple decision trees to make judgements about the sort of test of the test item. Just four test samples are falsely labeled in layer 1, and 2 test samples are incorrectly classified in two levels, out of around 4000 test data. The combination of DestinationIP and SourceIP, which has been determined as 0.25, is the most essential aspect of the random forests method in layer 1.

1.3 Discussion

It is noticeable that due of the utilization of a balanced and ensembled dataset, a significant improvement in the total confusion matrix has been observed. Balancing and ensemble approaches assist to manage constraints such as low available information, which ultimately helps to improve the model's predictive accuracy. The above study's assessment findings suggest that there is no such thing as an optimum machine learning technique for such challenges. Therefore, the packing strategy of the Random Forest model was chosen since it delivers more accuracy and consistency than some others do.

7 Conclusion and Future Work

Domain Name System (DNS) is one of the most essential Internet protocols that has been widely used since its inception. Many security flaws in the DNS protocol have been discovered over the years, prompting the development of extensions and other protocols to help make DNS more safe. One of these efforts is the DNS-over-HTTPS (DoH) protocol, which encrypts DNS packets using the HTTPS protocol, making DNS more private and addressing certain security problems. While DoH has been hailed for its simplicity of use and the security benefits it brings, it has not been thoroughly investigated to see how much it can assist with DNS's present vulnerabilities and what new vulnerabilities exist that must be addressed.
In the future I would like to do more in real network and instead of 20% I would like to do in 40 % with more number of training dataset that can be accomplished through data set. Also with less availablity of dataset I counld't do more as only 1 data set available to perform. With this DOH more traffic will genrateed and more data will transmit. In this thesis, I conducted a comprehensive examination of DNS security flaws and developed a taxonomy of potential DNS attacks. Using this taxonomy, I investigated DoH's effects on DNS security and examined the security elements of the DoH protocol. Secretive communications over DoH are one of the most serious security risks with the DoH protocol. DNS tunnels are types of DNS protocol connections that function by encapsulating data in DNS queries and answers.

8

Bibliography

[1]
The Economic Times, "Definition of Data Mining | What Is Data Mining ? Data Mining Meaning - the Economic Times," *The Economic Times*, 2019. https://economictimes.indiatimes.com/definition/data-mining
[2]
Y. M. Banadaki, "Detecting Malicious DNS over HTTPS Traffic in Domain Name System using Machine Learning Classifiers," *Journal of Computer Sciences and Applications*, vol. 8, no. 2, pp. 46–55, Aug. 2020, doi: 10.12691/jcsa-8-2-2.
[3]
"DoHBrw 2020 | Datasets | Research | Canadian Institute for Cybersecurity | UNB," *www.unb.ca*. https://www.unb.ca/cic/datasets/dohbrw-2020.html (accessed Mar. 06, 2021).

[4]
L. F. Gonzalez Casanova and P.-C. Lin, “Generalized Classification of DNS over HTTPS Traffic with Deep Learning,” *IEEE Xplore*, Dec. 01, 2021. https://ieeexplore.ieee.org/document/9689667 (accessed Aug. 11, 2022).

[5]
D. Vekshin, K. Hynek, and T. Cejka, “DoH Insight: Detecting DNS over HTTPS by Machine Learning,” doi: 10.1145/3407023.3409192.

[6]
M. Zago, M. Gil Pérez, and G. Martínez Pérez, “UMUDGA: A dataset for profiling DGA-based botnet,” *Computers & Security*, vol. 92, p. 101719, May 2020, doi: 10.1016/j.cose.2020.101719.

[7]
S. Ajmera and T. R. Pattanshetti, “A Survey Report on Identifying Different Machine Learning Algorithms in Detecting Domain Generation Algorithms within Enterprise Network,” *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, doi: 10.1109/icccnt49239.2020.9225357.

[8]
P. Yang, Y. Li, and Y. Zang, “Detecting DNS covert channels using stacking model,” *China Communications*, vol. 17, no. 10, pp. 183–194, Oct. 2020, doi: 10.23919/jcc.2020.10.013.

[9]
S. K. Singh and P. K. Roy, “Malicious traffic Detection of DNS over HTTPS using Ensemble Machine Learning,” *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 1061–1069, Mar. 2022, doi: 10.12785/ijcds/110185.

[10]
Y. Li, A. Dandoush, and J. Liu, “Evaluation and Optimization of learning-based DNS over HTTPS Traffic Classification,” *IEEE Xplore*, Oct. 01, 2021. https://ieeexplore.ieee.org/document/9615659 (accessed Aug. 15, 2022).

[11]
“Domain Name System (DNS) Security,” *compsec101.antibozo.net*. http://compsec101.antibozo.net/papers/dnssec/dnssec.html (accessed Aug. 15, 2022).

[12]
A. Ramdas and R. Muthukrishnan, “A Survey on DNS Security Issues and Mitigation Techniques,” *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, May 2019, doi: 10.1109/iccs45141.2019.9065354.

[13]
H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, “Intrusion detection system: A comprehensive review,” *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16–24, Jan. 2013, doi: 10.1016/j.jnca.2012.09.004.

[14]
M. Behnke *et al.*, “Feature Engineering and Machine Learning Model Comparison for Malicious Activity Detection in the DNS-Over-HTTPS Protocol,” *IEEE Access*, vol. 9, pp. 129902–129916, 2021, doi: 10.1109/ACCESS.2021.3113294.

[15]
C. Lu *et al.*, “An End-to-End, Large-Scale Measurement of DNS-over-Encryption: How Far Have We Come? CCS CONCEPTS • Networks → Application layer protocols; Network mea- surement; Naming and addressing. KEYWORDS Domane Name System, DNS Privacy, DNS-over-TLS, DNS-over- HTTPS, DNS Measurement ACM Reference Format,” 2019, doi: 10.1145/3355369.3355580.