Port Scanning Detection in Packet Capture (PCAP) File Using Decision Tree and Logistic Regression

Wang Mingye  
*WKWSCI*   
*Nanyang Technological University*Singapore  
mwang015@e.ntu.edu.sg

*Abstract*—In order to maintain the security of information systems, it is necessary for information system owners to detect malicious behaviors or cybercrime intrusions. However, some types of network malicious behaviors are hard to identify and monitor by humans or simple computer processes. Using a trained machine learning model will help us a lot since the machine learning model can judge whether a network behavior is malicious and can tell the type of this malicious behavior automatically with a network behavior record as the input. In this article, a decision tree model and a logistic regression model are separately trained on data extracted from a packet capture (PCAP) file, which contains the network behavior records of port scanning behavior and benign behaviors. The packet capture file is from CIC-IDS2017, an intrusion detection evaluation dataset. Both the trained models can tell whether a network behavior is a port scanning. The detailed processes of gathering the train data, preprocessing data and training models are included in this article. In theory, same methods can be applied to other PCAP files in CIC-IDS2017 that contain other types of malicious behavior, such as Brute Force, DDoS, XSS and Infiltration, but only Port Scanning dataset is used in this article due to the limitation of the author’s device hardware capability.

Keywords—Intrusion Detection, Machine Learning, Decision Tree, Logistic Regression, Port Scanning

# The forensic problem

Port scanning [1], a kind of malicious network behavior, is a method of determining which ports on a network are open and could be receiving or sending data. It is also a process for sending packets to specific ports on a host and analyzing responses to identify vulnerabilities. Port scanning can help to determine open or vulnerable server locations and diagnose security levels. Port scanning can also reveal the presence of security measures in place such as the firewall rule of the server.

This article aims to use machine learning models to find out the port scanning behaviors from a group of network behaviors whose information are extracted from a raw PCAP file. It is essentially a binary classification problem. That is, judging whether a network behavior is a port scanning or not.

# Survey of Related Work

## Methods to Conduct Intrusion Detection

As early as 1987, Denning, D. E. presented a model of a real-time intrusion-detection expert system capable of detecting break-ins, penetrations, and other forms of computer abuse is described. [2]

More recently, machine learning models are utilized to conduct intrusion detection. Almost all the machine learning algorithms that are able to do classification can be used for intrusion detection. Here are some examples: Liao, Y. & Vemuri, V. R. used KNN [3], Chen, W. H., Hsu, S. H., & Shen, H. P. used SVM [4] and Farnaaz, N., & Jabbar, M. A. used random forest [5].

Moreover, without doubt, deep learning models such as CNN and LSTM, which are also a kind of machine learning algorithms that are able to do classification, can be used to detect intrusion. Deep learning models can even directly take the binary content [6] of the entries in PCAP file as the input, rather than structured data, which release us from the feature engineering or data preprocessing.

Both the expert system and machine learning algorithms (deep learning is a subset of machine learning) are methods under artificial intelligence. Considering that intrusion detection in essence is a kind of prediction or judgement based on given features, which is exactly suitable to be solved via AI methods, no wonder most of published work used methods related to AI to conduct intrusion detection.

## Datasets of Intrusion Detection

Due to the volume of the internet nowadays, network intrusion behaviors can be ubiquitous. Therefore, there are many intrusion detection datasets whose data are collected from real life production network, though it is still possible to produce an artificial dataset via conducting specific attacks yourself to a network environment under your control.

Many intrusion datasets give structured or processed data rather than raw PCAP files. These data may be more similar to traditional machine learning or datamining datasets that are cell by cell, row by row and column by column. KDD Cup 1999 dataset [7] contains a wide variety of intrusions simulated in a military network environment. CTU-13 dataset [8] is a labeled dataset with botnet, normal and background traffic.

There are also datasets that are or includes PCAP files. ISCXIDS2012 [9] and CIC-IDS2017 [10], which are both from Canadian Institute of Cybersecurity, have the raw PCAP files that captured from network. Besides, CIC-IDS2017 has also the structured data that are concluded from PCAP files. CIC-IDS2017 is the dataset that this article uses, but I use the data that are newly extracted from PCAP files in my own way, instead of directly using the structured data given in the original dataset. Details of extracting data will be introduced in IV. Data Gathered.

# Slected Machine Learning Methods

## Decision Tree

Decision trees [11] are a non-parametric, which means the learning changes with the change of the fact (dataset) and has week prior, supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. As mentioned before, the question that this article solves is in essence a binary classification problem, so classification tree is used.

The learning result, or the classifier that the decision tree algorithm outputs is exact a tree-structured model. A decision tree contains a root node, several internal nodes and several nodes, where:

* Each non-leaf node represents a test on a feature attribute.
* Each branch represents the output of this feature attribute on a certain range.
* Each leaf node represents a specific category (classification result).

The process of using a decision tree to make decisions is to start from the root node, test the corresponding feature attributes in the items to be classified, and select the output branch according to its value until it reaches the leaf node. Finally, the category that the leaf node represents is the decision result.

The key step in constructing a decision tree is attributes splitting. Attributes splitting is to construct different branches at a certain node according to the different divisions of a feature, and the goal is to make the items that are classified in each split subset belong to the same category. There are some attribute splitting metrics, or splitting criterions: Information Gain (ID3 algorithm, Iterative Dichotomiser 3), Information Gain Ratio (C4.5 algorithm) and Gini Index (CART algorithm, Classification and Regression Tree).

In this article, the decision tree based on Gini index will be used. Here are the details about the splitting scheme about the tree on Gini index.

In the classification problem, assuming that there are K categories and the probability of the k-th category is pk, the expression of the Gini coefficient is:

Specifically, if the problem is binary classification problem, the expression of the Gini coefficient is:

For sample D, if D is divided into two parts D1 and D2 according to a certain value of feature A, then under the condition of feature A, the Gini index expression of D is:

For classification tree, the CART algorithm chooses the splitting scheme where the attribute minimizes child nodes’ Gini Index. That is, to choose an attribute as the splitting node that minimizes:

## Logistic Regression

Logistic regression [12], despite its name, is a linear model for classification rather than regression. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. The essence of logistic regression is to assume that the data follows this distribution, and then use maximum likelihood estimation to estimate parameters.

When doing classification, the expected result should be a probability or a set of probabilities which indicates how possible the sample belongs to a specific category. For probability, it has a range of [0, 1]. Normally, a linear regression cannot constantly output a value within [0, 1], but we can put the linear regression expression into another function and now the combined function can always output [0, 1]. In logistic regression, the ‘another’ function is sigmoid function:

Let p mean the output probability. Substitute the linear regression expression into t:

Let’s y is the result of a binary classification that 1 means one category and 0 means another, then:

Now the principle formula of how logistic regression predicts is clarified. The question is, with given a sample dataset x and its corresponding classification results y, how to find parameters , which makes logistic regression can well fit the data.

Usually, to train a machine learning model and make it fit the data, the aim is to minimize the loss function. In logistic regression, the loss function is cross-entropy function:

And the cross-entropy function without is exactly the likelihood function of logistic regression model. That is, minimizing the loss function is equivalent to maximizing the likelihood function. Stochastic gradient descent method and newton’s method can be used to find the solution. Besides, to avoid overfitting, regularization can be added to logistic regression. In this article, regularization is used.

# Data Gathering

As mentioned before, dataset CIC-IDS2017 [10] is used in this article and only port scanning is concerned. In this dataset, port scanning happened at afternoon, Friday, July 7, 2017, so I only downloaded the raw PCAP file named Friday-WorkingHours.pcap. This single file’s size has reached 8.2G. And total PCAP files (5 days record) in this dataset have the size of more than 50G. The too large size in total, which leads to difficulties in downloading, storing in the disk and reading by program, is one reason that I cannot consider all malicious behavior types but only port scanning in this dataset.

In addition, the structured data in CIC-IDS2017, which are concluded from PCAP files, are also needed, because the structured data are labeled while the raw PCAP file have no labels. We cannot know which behavior record in the PCAP file is port scanning behavior without the structured data’s labels. The file name of the needed structured data in dataset is GeneratedLabelledFlows.zip.

## Confirm Port Scanning PCAP Records and Filter Raw PCAP with Time Range

To train a classification model, we must use supervised learning method on labeled data. Therefore, we cannot directly train the model based on unlabeled raw PCAP data. Here I read the structured data of port scanning records (Friday-WorkingHours-Afternoon-PortScan.pcap\_ISCX.csv) given in CIC-IDS2017 and found that the source IP of all entries (rows) which are labeled with ‘PortScan’ is 172.16.0.1. And the rest are all normal behaviors entries and are labeled with ‘Benign’. Here is what the first entry that is a port scanning behavior looks like in the csv file (some of the fields):

1. The First ‘PortScan’ Entry in Structured Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source IP | … | Timestamp | … | Label |
| 172.16.0.1 | … | 7/7/2017 1:05 | … | PortScan |

Moreover, the timestamp of the last entry that is a port scanning behavior is ‘7/7/2017 3:23’. We can conclude that the port scanning behaviors are all within ‘7/7/2017 1:05’ and ‘7/7/2017 3:23’. To reduce the size of data, I decided to filter the raw PCAP file with the time range 7/7/2017 1:05 to 7/7/2017 3:23. This also reduces the interference of excessive ‘Benign’ data. Even in the filtered and processed data, there are 1487956 ‘Benign’ entries and only 162657 ‘PortScan’ entries.

However, the time range 7/7/2017 1:05 to 7/7/2017 3:23 cannot be directly used to filter the raw PCAP data. First, as stated, this is an ‘afternoon’ data, so the time in 24-hour clock should be 13:05 and 15:23. Second, the timestamp in the structured data is in Canada time zone but the time display format in Wireshark can only be set to UTC time or local time (UTC+8 in Singapore).

To find out the correct time range that can be applied to filter raw PCAP data in Wireshark, here I set Wireshark time display format to UTC time:

图形用户界面, 文本, 应用程序, 聊天或短信

描述已自动生成

1. Set Wireshark Time Display Format to UTC Time

Then I found that Canadian time can be from UTC-3 to UTC-7 in summertime (because the date 7/7 is in the summertime period), so I decided to use Canada time 13:05 to compute the corresponding UTC time, from the difference 3 to the difference 7, and compare the very first PCAP entry whose Source IP is 172.16.0.1 with the first ‘PortScan’ entry in structured data to see whether they are consistent.

Finally, I confirmed that the Canada time 13:05 in the structured data is in time zone UTC-3, which means the time range that should be applied to Wireshark to filter the PCAP data in UTC time is 16:05 to 18:23. Specifically, the time of the very first entry where 172.16.0.1 appears as the Source IP is 16:05:33.754382.

I applied the time range 16:05:33.754382 to 18:24 to the raw PCAP file, Friday-WorkingHours.pcap (8.2G) and save the filtered PCAP entries as another file, whose size is only 900M. The condition expression of this time range in Wireshark filter is “frame.time >= "2017-07-07 16:05:33.754382" && frame.time <= "2017-07-07 18:24:00.0"”.

表格

描述已自动生成

1. Apply the Time Range Condition to Filter PCAP Data

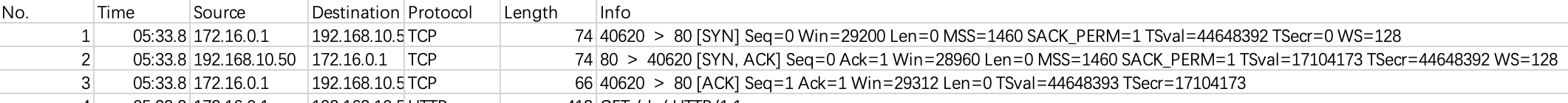
## Generate Structured Data (.csv) from PCAP Data (.pcap)

So far, I have gathered a more accurate raw PCAP file than the original one, but I still have to get structured data to train machine learning models, decision tree and logistic regression. Initially, I try to use Wireshark to open the PCAP file and export the packet dissections as a .csv file. However, the data in exported .csv file are not structured and detailed enough. It leaves unstructured information in the text of ‘Info’ column and has not enough details of the packet.

图形用户界面, 文本, 应用程序, 聊天或短信

描述已自动生成

1. Use Wireshark to Export Packet Dissections as CSV File



1. The Example of the Exported CSV Data by Wireshark

According to Shanto Roy’s blog [13], the right way to gather the detailed and structured CSV data from PCAP file is using ‘tshark’ command. By defining the needed fields in ‘tshark’ command, one can gather whatever packets’ information he/she wants from the PCAP file.

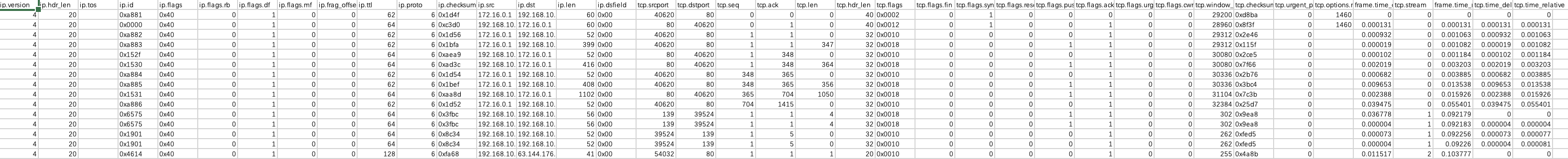
Usually, ‘tshark’ tool is included in the installed Wireshark. As a Mac user, first I set a soft link, in order to directly use ‘tshark’ in Terminal:

ln -s /Applications/Wireshark.app/Contents/MacOS/tshark /usr/local/bin/tshark

Initially, I had no idea with which kinds of information are exactly needed to train an accurate model, so I referred to Shanto Roy’s practice as well as Tao Liu’s practice and I extracted all the fields used by both. Here is my ‘tshark’ command to extract data from previous PCAP file (the previous PCAP file is named PortScan\_Begin\_to\_End.pcap and the CSV file to generate is named PortScan\_Begin\_to\_End\_tshark\_deltail.csv):

tshark -r PortScan\_Begin\_to\_End.pcap -T fields -E header=y -E separator=, -E quote=d -E occurrence=f -e ip.version -e ip.hdr\_len -e ip.tos -e ip.id -e ip.flags -e ip.flags.rb -e ip.flags.df -e ip.flags.mf -e ip.frag\_offset -e ip.ttl -e ip.proto -e ip.checksum -e ip.src -e ip.dst -e ip.len -e ip.dsfield -e tcp.srcport -e tcp.dstport -e tcp.seq -e tcp.ack -e tcp.len -e tcp.hdr\_len -e tcp.flags -e tcp.flags.fin -e tcp.flags.syn -e tcp.flags.reset -e tcp.flags.push -e tcp.flags.ack -e tcp.flags.urg -e tcp.flags.cwr -e tcp.window\_size -e tcp.checksum -e tcp.urgent\_pointer -e tcp.options.mss\_val -e frame.time\_delta -e tcp.stream -e frame.time\_relative -e tcp.time\_delta -e tcp.time\_relative > PortScan\_Begin\_to\_End\_tshark\_deltail.csv

After running this command in Terminal, I got the .csv file with detailed and structured data in it. Here the figure may be not clear enough. Compared to the data exported via Wireshark that only has 5 fields (without No.), the data exported by ‘tshark’ command has 39 fields.



1. The Example of the Exported CSV Data by ‘tshark’ Command

# Data Preprocessing

Now I have a new dataset in .csv format, which is much more suitable to be used to train machine learning model than other data before. However, the preprocessing of current data is still necessary before I use it to train models.

The full code of preprocessing is in Appendix – B. preprocess\_pcap\_csv.py.

First, use pandas.read\_csv() to read the .csv file as a pandas.DataFrame in program. Then I check the data type (pandas.DataFrame.info())and the situation of the missing value of each field (pandas.DataFrame.isnull().sum()). Here are the results:

1. Field Name, Data Type and Total Missing Valuue Amount

|  |  |  |
| --- | --- | --- |
| Field | Data Type | Total Missing Value Amount |
| ip.version | float64 | 14792 |
| ip.hdr\_len | float64 | 21690 |
| ip.tos | float64 | 1885472 |
| ip.id | object | 21690 |
| ip.flags | object | 21690 |
| ip.flags.rb | float64 | 21690 |
| ip.flags.df | float64 | 21690 |
| ip.flags.mf | float64 | 21690 |
| ip.frag\_offset | float64 | 21690 |
| ip.ttl | float64 | 21690 |
| ip.proto | float64 | 21690 |
| ip.checksum | object | 21690 |
| ip.src | object | 21690 |
| ip.dst | object | 21690 |
| ip.len | float64 | 21690 |
| ip.dsfield | object | 21690 |
| tcp.srcport | float64 | 234853 |
| tcp.dstport | float64 | 234853 |
| tcp.seq | float64 | 234853 |
| tcp.ack | float64 | 234859 |
| tcp.len | float64 | 234989 |
| tcp.hdr\_len | float64 | 234859 |
| tcp.flags | object | 234859 |
| tcp.flags.fin | float64 | 234859 |
| tcp.flags.syn | float64 | 234859 |
| tcp.flags.reset | float64 | 234859 |
| tcp.flags.push | float64 | 234859 |
| tcp.flags.ack | float64 | 234859 |
| tcp.flags.urg | float64 | 234859 |
| tcp.flags.cwr | float64 | 234859 |
| tcp.window\_size | float64 | 234859 |
| tcp.checksum | object | 234859 |
| tcp.urgent\_pointer | float64 | 234859 |
| tcp.options.mss\_val | float64 | 1663957 |
| frame.time\_delta | float64 | 0 |
| tcp.stream | float64 | 234859 |
| frame.time\_relative | float64 | 0 |
| tcp.time\_delta | float64 | 234859 |
| tcp.time\_relative | float64 | 234859 |

Next, I did the following operations:

### pandas.DataFrame.drop\_duplicates(): Drop the rows that are identical. Only keep one of them.

### pandas.DataFrame.dropna(axis = 0, subset = ["ip.src"]): Drop the rows whose ‘ip.src’ field is missing.

### pandas.DataFrame.dropna(axis = 0, subset = ["tcp.window\_size"]): Drop the rows whose ‘ip.src’ field is missing.

### pandas.DataFrame.drop(columns=["ip.tos"]): Drop the column of field ‘ip.tos’, whose values are all mssing.

### pandas.DataFrame.fillna(0, inplace=True): Fill all cells that misses value with 0.

### pandas.DataFrame.applymap(hexadecimal\_to\_decimal): Change hexadecimal number to decimal number. The fields whose value is hexadecimal number are: ‘ip.id’, ‘ip.checksum’, ‘ip.dsfield’, ‘tcp.flags’ and ‘tcp.checksum’. The specific hexadecimal\_to\_decimal funcrtion can be found in Appendix – B. preprocess\_pcap\_csv.py.

### pandas.DataFrame.apply(lambda x: 1 if x["ip.src"] == "172.16.0.1" else 0, axis=1): Add the ‘label’ column. The row whose ‘ip.src’ value is ‘172.16.0.1’ will be assigned the label 1, which means this row represents a port scanning behavior, as illustrated in VI. Data Gathering – A. Conirm Port Scanning PCAP Records and Filter Raw PCAP with Time Range. Otherwise, the label is 0.

### pandas.DataFrame.to\_csv(): Save the processed data as a new file.

# Model Training

First,

#### Selection: Highlight all author and affiliation lines.

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya | aaa |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. Avast Business. (n.d.). What is port scanning and how does it work? | Avast. Avast Antivirus. Retrieved March 16, 2022, from https://www.avast.com/en-sg/business/resources/what-is-port-scanning.
2. Denning, D. E. (1987). An intrusion-detection model. IEEE Transactions on software engineering, (2), 222-232.
3. Liao, Y., & Vemuri, V. R. (2002). Use of k-nearest neighbor classifier for intrusion detection. Computers & security, 21(5), 439-448.
4. Chen, W. H., Hsu, S. H., & Shen, H. P. (2005). Application of SVM and ANN for intrusion detection. Computers & Operations Research, 32(10), 2617-2634.
5. Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.
6. A.I.L. (2021, September 16). Detection of abnormal network traffic through deep learning. CSDN. Retrieved March 16, 2022, from https://blog.csdn.net/weixin\_39032619/article/details/120337693
7. KDD Cup. (1999, October 28). KDD Cup 1999 Data. KDD. Retrieved March 17, 2022, from http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
8. Stratosphere Laboratory. (n.d.). The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background traffic. Stratosphere IPS. Retrieved March 17, 2022, from https://www.stratosphereips.org/datasets-ctu13/
9. Canadian Institute for Cybersecurity. (2012, May). IDS 2012 | Datasets | Research | Canadian Institute for Cybersecurity | UNB. Retrieved March 18, 2022, from https://www.unb.ca/cic/datasets/ids.html
10. Canadian Institute for Cybersecurity. (2017, July). IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB. Retrieved March 18, 2022, from https://www.unb.ca/cic/datasets/ids-2017.html
11. Scikit-learn Developers. (n.d.). 1.10. Decision Trees. Scikit-Learn. Retrieved March 18, 2022, from https://scikit-learn.org/stable/modules/tree.html#tree
12. Scikit-learn Developers. (n.d.). 1.1. Linear Models. Scikit-Learn. Retrieved March 20, 2022, from https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression
13. Roy, S. (2021, February 5). *Convert PCAP files to CSV for Network Traffic Analysis*. Roy’s Blog. Retrieved March 23, 2022, from https://shantoroy.com/networking/convert-pcap-to-csv-using-tshark/
14. Liu, T. (2020, June 1). *Build a Machine Learning Model for Network Flow Classification*. LinkedIn. Retrieved March 23, 2022, from https://www.linkedin.com/pulse/build-machine-learning-model-network-flow-tao-liu/

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**