Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 3, No. 1, Mei 2025, hlm 13-23

On The Comprehensive Analyses of CTU-13 Botnet Dataset for Cyber Security Researches

Jimoh Rasheed Gbenga[1], Oyelakin Akinyemi Moruff[2]
[1]Department of Computer Science, Faculty of Communication and Information Sciences, University of Ilorin, Nigeria
[2]Department of Computer Science, College of Information and Communication Technology, Crescent University, Abeokuta, Nigeria
moruff.oyelakin@cuab.edu.ng[1]*

Correzponding Author: *moruff.oyelakin@cuab.edu.ng

ABSTRACT

Attackers use malware to launch attacks in the internet and corporate networks. Over the years, machine learning techniques have been found promising for the classification of these attacks because they have the ability to identify unknown threats. Botnets are networks of compromised devices and have been found to be powerful threat vectors that are used against modern systems because they use command and control (C2) characteristics which make their detection very difficult. Generally, to build attack detection models, intrusion datasets are employed. Comprehensive study of the benchmarking datasets used in intrusion detection researches can provide different actionable insights to other researchers. There have been studies that investigated the analyses of datasets for building intrusion detection systems. However, there has been less focus on the analysis of intrusion detection datasets that are used specifically for botnets detection. This study reported an overview of a popular botnet dataset named CTU-13. Thereafter, the work carried out detailed exploratory analysis of the dataset. The study equally sought to identify if the dataset is representative enough for Machine Learning based botnet detection studies. All the thirteen scenarios in the dataset were used for the experimentations. The exploratory analyses were carried out on each of the thirteen scenarios of the dataset with a view to gaining better understanding of the patterns and characteristics of data in each of them. The information obtained from the overview and exploratory analyses provided actionable insights on how to better use the datasets for improved botnet classification. The challenges of using the captures of the dataset were also identified. In particular, the exploratory investigation of the thirteen captures of the CTU-13 dataset revealed that it has very complex patterns, contain mixed data types and suffers from high class imbalance problem. The results of the exploratory analyses can guide the decision of future cyber security researches. Thus, improved machine learning-based botnet detection models can be built by attending to the issues in the dataset.

ARTICLE INFO

Article history
Received: 31 Desember 2024
Revised: 30 April 2025
Accepted: 10 Mei 2025

Keywords
Botnet Classification,
Exploratory Data Analysis,
Machine Learning Methods,
Dataset Imbalance

1. Introduction

Malware of different types are used to launch attacks in the cyber space and enterprise networks. Botnets are used to launch attacks through the Command and Control (C&C) server [1]. Over the years, there have been different approaches used for the botnet detection and mitigation. One of the common approaches for classifying network attacks is the use of machine learning (ML)

algorithms ([2]; [3]). Statistical learning and machine learning classification approaches have been found to be very promising in different areas of network or cyber security. However, machine learning-based botnet detection studies have suffered from the dearth of large and representative datasets [4]. This is further confirmed by researchers in [5] who argued that the two major difficulties with botnet detection researches are the lack of big and rich datasets and that the few available ones are either too small and or are synthetically generated.

Intrusion Detection Systems that are based on ML methods have been found to be effective and accurate in detecting unknown networks attacks [6] when compared to signature-based techniques. In order to step up efforts against intrusions, benchmark datasets that are used have been developed over times [7]. As part of these efforts against which is a threat in network and internet space, some publicly available botnet datasets have been released for researches. Some examples of such datasets are: ISOT botnet dataset by [8], ISCX botnet dataset by [9], CTU-13 dataset by [4] and many others. Authors in [10] have pointed out that data representativeness is very important when it comes to drawing inference from data through machine learning models. The need for this is to ensure that such models are not bias and are fair when fed with input data. This study focuses on exploring CTU-13 dataset. Preliminary investigation equally showed that past studies on exploratory analyses of datasets are based on general intrusion detection datasets. To the best of our knowledge, there is less focus on analysis of the chosen botnet dataset.This work carried out an overview and exploratory analyses of all the thirteen captures in the botnet dataset. The purpose of the exploratory analyses of the said dataset is to provide actionable insights on how to use the datasets for improved botnet detection purposes by other researchers.

2. Literature Review

Researchers in [11] built Tree-Based Learning Models for Botnet Malware Classification using three different selected sub-samples of the CTU-13 dataset. The authors argued that the sub-sample were carried out randomly so has not alter the distributions of the patterns in the dataset. They further claimed that promising results were achieved by the classification models built for the botnet evidence. Similarly, authors in [12] carried out a study that reported an overview and exploratory analyses of CICIDS 2017 Intrusion Detection Dataset. The study provided actionable insights for security researchers who may be interested in using the dataset for bench-marking machine learning models. The work did not made mention of how representative the data is.

Researchers in [7] performed a detailed analysis of eight different datasets that are used for Intrusion Detection Systems (IDSs). The paper introduced a detailed analysis of benchmark datasets for Network Intrusion Detection Systems. The datasets covered include: KDD99, NSL-KDD, KYOTO 2006+, ISCX2012, UNSW-NB 15, CIDDS-001, CICIDS2017, and CSE-CIC-IDS2018. Furthermore, authors in [13] carried out exploratory analysis of the ISOT Cloud Intrusion Dataset popularly called ISOT-CID. The focus of the work was to explore cloud anomaly detection with the use of three different machine learning techniques. It was reported that the dataset contains various attacks and normal activities gathered in a real cloud environment and promising for intrusion detection studies.

[4] performed a detailed analysis of a recent intrusion detection-based dataset named CICIDS2017. During the analysis, the authors argued that some of the drawbacks of the dataset were found. Furthermore, the authors proposed to fix the identified problems and produce a version of the dataset called optimized CICIDS2017 dataset. It was argued that best results were obtained from the models built in the optimized dataset based on the metrics used for the evaluation. Authors in [15] equally proposed an analysis of a dataset named CICIDS2017 dataset that was released for Intrusion Detection System studies. The focus of the work is the exploration of the characteristics of the dataset. Thereafter, the characteristics of the datasets were outlined. The study presented a combined dataset by eliminating such issues for better classification and detection of any future intrusion detection engine. Also, authors in [8] developed a model for the detection of HTTP Botnet based on DNS Traffic Analysis and Application profiling. The emphasis of the work is on identifying HTTP-based botnet that tries to bury illegitimate DNS traffic in the legitimate ones. Going by this approach, it is assumed that having a large dataset containing different traffic and attack types like the ones in CTU-13 dataset may be the best in any botnet study.

3. Research Method

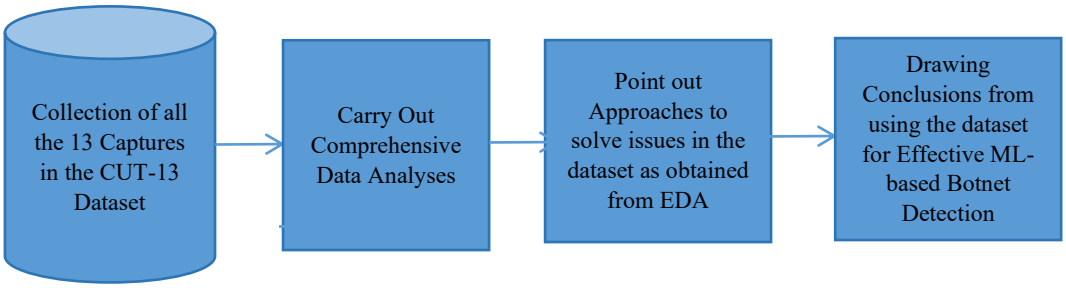

Figure 1. Methodological Flow in the Study

The method used in this study is as summarily captured in figure 1. The work focuses on detailed exploratory analyses of the collected CTU-13 dataset as an example of all-in-all botnet dataset. The researchers first of all collected some relevant articles on botnet detection approaches and provided overview on their strengths and limitations. This work specifically identified the need to investigate a new dataset named CTU-13 as released by authors in [4] and thus collected it in whole. The netflow portion of the dataset was used in all the experimental analyses having save them in csv format. Thereafter, exploratory analyses was carried out on all the thirteen scenarios of the dataset. The dataset was chosen because it contains real-world traces and is a good representative of datasets for botnet detection studies as argued by [4]. All the experimentations were carried out in a Python IDE environment named Spyder. The researchers equally sought to identify if the dataset is representative enough for Machine Learning based botnet detection studies based on the various types of exploratory analyses carried out.

4. Findings

The findings of the study are grouped into two. The first being an overview of botnets. The second part is on results are based on the detailed exploratory analyses.It is argued herein that the analyses will provide a good ground for future machine learning-based botnet detection studies.

4.1 Overview of the CTU-13 Botnet Dataset

The dataset used in this study is called CTU-13 dataset. It is a very large real-life dataset that contains several botnet samples and millions of instances.The dataset that is grouped into thirteen different scenarios and can be downloaded as a whole from the following link: https://mcfp.felk.cvut.cz/publicDatasets/CTU-13-Dataset/CTU-13-Dataset.tar.bz2. Authors of the dataset in [4] mentioned that on each scenario of the dataset there is a specific malware which used several protocols and performed different actions. Table 1 shows the characteristics of the botnet scenarios.

Table 1. Characteristics of the botnet scenarios in CTU-13 dataset [4]

ID	IRC	SPAM	CF	PS	DDoS	FF	P2P	US	HTTP	Note (Additional)
1	Yes	Yes	Yes							
2	Yes	Yes	Yes							
3	Yes			Yes				Yes		
4	Yes							Yes		UDP and ICMP DDoS
5		Yes		Yes					Yes	Scan Web Proxies
6				Yes						Proprietary C&C RDP
7									Yes	Chinese hosts
8				Yes						Proprietary C&C, Net-BIOS, STUN.
9	Yes	Yes	Yes	Yes	Yes					
10	Yes				Yes			Yes		UDP DDoS
11	Yes							Yes		ICMP DDoS
12							Yes			Synchronization
13		Yes							Yes	Captcha, Web Mail

Table 1 shows the characteristics of the botnet scenarios in CTU-13 dataset as released by authors in [4]. A preliminary analysis of the values in table 1 revealed that the CTU-13 dataset contains different botnet samples with varying protocol characteristics and attacks. It can be further deduced that the major attacks include SPAM and DDoS attacks. Interestingly, to understand the dataset better, the number of records in each captures of the dataset is as shown in table 2. The table equally provides the number of traffic in the dataset, both both in pcap and netflow formats. It equally shows the samples of the botnet in each of the dataset scenario.

Table 2: Amount of Data in 13 captures of CTU-13 dataset [4]

Id	Duration(hrs)	No of Packets	No of NetFlows	Size	Bot	No of Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menu	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

Looking at the number of packets and netflows in the dataset as shown in table 2, it can deduced that to train a ML model for the classification of botnets in each of the scenarios (captures) requires some level of innovation so as to ensure that the models built are computationally less expensive. The thirteen captures are of different sizes and patterns.

4.2 Results of Exploratory Analyses of CTU-13 Datasets

Based on the exploratory analysis of the CTU-13 Dataset, it was observed that each of the captures in the dataset contains fourteen input features with hundreds of thousands of instances and millions in some cases. The input features in dataset include: StartTime, Duration, Protocol, SrcAddr, SrcPort, Direction, DstAddr, Destination port, State of the transmission, sTos, dTos, TotPkts, Totbytes, SrcBytes. The target feature is named label which is multi-class in nature. As shown in table 3, it was also observed that the dataset contains mixed data types. Thus, it will be required that any researcher using this dataset to build botnet detection models address the multi-class, mixed data type and big data issues in the dataset. The mixed data types are found in the dataset as shown in Table 3.

Table 3. Categorical and Non-categorical data types in the CTU-13 Dataset

Feature	Feature Description	Feature Data Type
StartTime	Start Time of the Netflow	Object
Dur	Duration of the flow	float64
Proto	Protocol	Object
SrcAddr	Source Address	Object
Sport	Source port	Object
Dir	Traffic Direction	Object
DstAddr	Destination Address	Object
Dport	Destination Port	Object
State	State of the flow	Object
sTos	Type of Service from service to source	float64
dTos	Type of Service from destination to source	float64
TotPkts	Total Packets	int64
TotBytes	Total Bytes	int64
SrcBytes	Source Bytes	int64

Table 3 is an outcome of EDA showing the different data types in each of the thirteen captures of the dataset. The values in table 3 are as obtained from the exploratory analyses.

4.2.1 Summary Statistics of the Pre-processed Features in the Datasets

Furthermore, the exploratory analysis is used to know the statistical summary of both the numerical and categorical features. The categorical features have been pre-processed using label encoding before the statistical summary was computed. Tables so so

Table 4. Summary Statistics of Scenario 1 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
Count	1.048575e+06	1.048575e+06	...	1.048575e+06	1.048575e+06
Mean	5.243007e+05	4.929120e+02	...	2.382442e+03	4.202135e+01
Std	3.027273e+05	1.061373e+03	...	3.628309e+05	4.975838e+01
Min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	2.621435e+05	3.070000e-04	...	7.900000e+01	7.000000e+00
50%	5.242870e+05	1.319000e-03	...	8.300000e+01	7.000000e+00
75%	7.864305e+05	1.351720e+01	...	2.380000e+02	1.110000e+02
Max	1.442420e+06	3.600000e+03	...	2.484051e+08	1.120000e+02

Table 5. Summary Statistics of Scenario 2 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	1.808122e+06	1.808122e+06	...	1.808122e+06	1.808122e+06
mean	9.040600e+05	4.006726e+02	...	2.210684e+03	5.270104e+01
Std	5.219596e+05	9.516550e+02	...	3.034949e+05	6.034864e+01
Min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	4.520302e+05	3.140000e-04	...	7.900000e+01	5.000000e+00
50%	9.040595e+05	2.147000e-03	...	8.500000e+01	6.000000e+00
75%	1.356090e+06	8.996974e+00	...	3.100000e+02	1.300000e+02
Max	1.808120e+06	3.600034e+03	...	2.485222e+08	1.310000e+02

Table 6. Summary Statistics of Scenario 3 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	4.710638e+06	4.710638e+06	...	4.710638e+06	4.710638e+06
mean	2.355311e+06	1.779531e+02	...	6.983991e+03	2.640858e+01
std	1.359841e+06	6.783478e+02	...	2.239477e+06	2.208885e+01
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	1.177656e+06	2.410000e-04	...	7.600000e+01	4.000000e+00
50%	2.355312e+06	3.620000e-04	...	8.100000e+01	3.000000e+01
75%	3.532969e+06	3.096770e-01	...	2.220000e+02	4.900000e+01
max	4.710623e+06	3.600000e+03	...	3.423408e+09	5.000000e+01

Table 7. Summary Statistics of Scenario 4 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	1.121076e+06	1.121076e+06	...	1.121076e+06	1.121076e+06
mean	5.605360e+05	2.313872e+02	...	4.946739e+03	2.823176e+01
std	3.236262e+05	7.434456e+02	...	9.542122e+05	2.547417e+01
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	2.802678e+05	2.870000e-04	...	7.800000e+01	4.000000e+00
50%	5.605355e+05	7.610000e-04	...	8.500000e+01	6.000000e+00
75%	8.408042e+05	2.240842e+00	...	4.660000e+02	5.600000e+01
max	1.121073e+06	3.657061e+03	...	9.042148e+08	5.700000e+01

Table 8. Summary Statistics of Scenario 5 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	129832.000000	129832.000000	...	1.298320e+05	129832.000000
mean	64915.500000	77.439645	...	5.081775e+03	35.524801
std	37479.414412	282.290550	...	5.179676e+05	34.369575
min	0.000000	0.000000	...	0.000000e+00	0.000000
25%	32457.750000	0.000304	...	7.600000e+01	5.000000
50%	64915.500000	0.000739	...	8.100000e+01	6.000000
75%	97373.250000	0.643476	...	4.600000e+02	75.000000
max	129831.000000	1805.828491	...	1.365468e+08	76.000000

Table 9. Summary Statistics of Scenario 6 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	558919.00000	558919.000000	...	5.589190e+05	558919.000000
mean	279459.00000	244.257943	...	1.564192e+04	26.796904
std	161346.16189	762.289286	...	5.536115e+05	21.934023
min	0.00000	0.000000	...	0.000000e+00	0.000000
25%	139729.50000	0.000290	...	7.800000e+01	5.000000
50%	279459.00000	0.000681	...	8.300000e+01	30.000000
75%	419188.50000	2.237202	...	4.660000e+02	49.000000
max	558918.00000	3600.000000	...	2.517715e+08	50.000000

Table 10. Summary Statistics of Scenario 7 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	114077.000000	114077.000000	...	1.140770e+05	114077.000000
mean	57038.000000	84.655646	...	9.048453e+03	24.116728
std	32931.337666	254.416485	...	2.842253e+05	22.304197
min	0.000000	0.000000	...	0.000000e+00	0.000000
25%	28519.000000	0.000319	...	7.700000e+01	5.000000
50%	57038.000000	0.001729	...	8.100000e+01	6.000000
75%	85557.000000	1.376178	...	4.320000e+02	50.000000
max	114076.000000	1277.465088	...	4.074933e+07	51.000000

Table 11. Summary Statistics of Scenario 8 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	2.954230e+06	2.954230e+06	...	2.954230e+06	2.954230e+06
mean	1.477113e+06	3.043375e+02	...	9.778871e+03	3.052950e+01
std	8.528114e+05	8.556113e+02	...	2.014241e+06	2.626169e+01
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	7.385572e+05	2.740000e-04	...	7.800000e+01	6.000000e+00
50%	1.477114e+06	4.820000e-04	...	8.300000e+01	6.000000e+00
75%	2.215669e+06	6.789455e-01	...	2.840000e+02	5.800000e+01
max	2.954225e+06	3.600000e+03	...	2.692621e+09	5.900000e+01

Table 12. Summary Statistics of Scenario 9 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	2.087508e+06	2.087508e+06	...	2.087508e+06	2.087508e+06
mean	1.043748e+06	2.945965e+02	...	7.418329e+03	4.046411e+02
std	6.026107e+05	8.375559e+02	...	1.647297e+06	4.334299e+02
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	5.218728e+05	3.200000e-04	...	7.700000e+01	6.000000e+00
50%	1.043748e+06	9.890000e-04	...	8.300000e+01	6.000000e+00
75%	1.565624e+06	5.064932e+00	...	2.780000e+02	9.060000e+02
max	2.087501e+06	3.600080e+03	...	2.133291e+09	9.070000e+02

Table 13. Summary Statistics of Scenario 10 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	1.309791e+06	1.309791e+06	...	1.309791e+06	1.309791e+06
mean	6.548938e+05	2.538714e+02	...	8.064216e+03	5.010197e+01
std	3.781034e+05	7.694298e+02	...	1.253815e+06	4.972972e+01
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	3.274475e+05	2.610000e-04	...	7.900000e+01	4.000000e+00
50%	6.548930e+05	8.780000e-04	...	9.000000e+01	6.000000e+00
75%	9.823405e+05	2.277852e+00	...	9.290000e+02	1.090000e+02
max	1.309788e+06	3.600019e+03	...	1.233900e+09	1.100000e+02

Table 14. Summary Statistics of Scenario 11 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	107251.000000	107251.000000	...	1.072510e+05	107251.000000
mean	53625.000000	49.906932	...	3.251469e+03	22.829633
std	30960.841198	169.009213	...	1.503550e+05	23.203360
min	0.000000	0.000000	...	0.000000e+00	0.000000
25%	26812.500000	0.000293	...	7.700000e+01	6.000000
50%	53625.000000	0.000917	...	8.500000e+01	6.000000
75%	80437.500000	0.348301	...	5.300000e+02	56.000000

	StartTime	Dur	...	SrcBytes	Label
Max	107250.000000	971.288147	...	2.287287e+07	57.000000

Table 15. Summary Statistics of Scenario 12 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	325471.000000	325471.000000	...	3.254710e+05	325471.000000
mean	162733.868557	216.158316	...	6.246470e+03	32.894897
Std	93955.037167	707.696942	...	7.979330e+05	32.231373
Min	0.000000	0.000000	...	0.000000e+00	0.000000
25%	81366.500000	0.000281	...	7.900000e+01	4.000000
50%	162734.000000	0.000998	...	8.500000e+01	6.000000
75%	244101.500000	1.922384	...	4.320000e+02	70.000000
max	325468.000000	3600.000000	...	3.452777e+08	71.000000

Table 16. Summary Statistics of Scenario 13 of the Pre-processed CTU-13 Dataset

	StartTime	Dur	...	SrcBytes	Label
count	1.925149e+06	1.925149e+06	...	1.925149e+06	1.925149e+06
mean	9.625733e+05	3.276373e+02	...	3.441054e+03	5.089235e+01
std	5.557420e+05	8.879237e+02	...	6.720011e+05	5.279416e+01
min	0.000000e+00	0.000000e+00	...	0.000000e+00	0.000000e+00
25%	4.812870e+05	2.720000e-04	...	7.700000e+01	6.000000e+00
50%	9.625740e+05	5.130000e-04	...	8.300000e+01	6.000000e+00
75%	1.443859e+06	1.732091e+00	...	2.620000e+02	1.130000e+02
max	1.925146e+06	3.600035e+03	...	5.055238e+08	1.140000e+02

4.2.2 Results of the missing values handling

Table 17. Summary of instances in CTU-13 Dataset before and after missing values handling

Scenario 1	No of Input features	No of Target Feature	No of original samples	Total number of missing values	No of reduced samples after deletion
1	14	1	2,824,636	205,296	2,619,340
2	14	1	1,808,122	273,815	1,534,307
3	14	1	4,710,638	534,903	4,166,735
4	14	1	1,121,076	89,301	1,031,775
5	14	1	129,832	7,683	122,149
6	14	1	558,919	37,729	521,190
7	14	1	114,077	7,637	106,440
8	14	1	2,954,230	185,759	2,768,471
9	14	1	2,087,508	181,829	1,905,679
10	14	1	1,309,791	199,261	1,110,530
11	14	1	107,251	17,833	89,368
12	14	1	325,471	30,201	295,270
13	14	1	1,925,149	147,586	1,777,563

Table 17 showed the results of the instances before and after handling the missing values in each of the thirteen captures of the CTU-13dataset. The last column in Table 17 can be obtained when deletion strategy is used for handling the missing values in the dataset.

Table 18. Key Summary of the 13 captures in the CTU-13 Dataset

S/N	Dataset Name	Author and Year	Brief Information about the Dataset	Class	Data Types	Attacks Types	Available Formats	Missing Values	Partial or Total Botnet Samples
1	CTU-13 Dataset	Garcia et al., 2014 [4]	The dataset is a complete dataset on botnets. It is a real-life botnet dataset that is contained in thirteen different captures	Multi Class	Mixed	Several	Netflow and PCAPs	YES	The samples and attacks are totally botnet based

S/N	Dataset Name	Author and Year	Brief Information about the Dataset	Class	Data Types	Attacks Types	Available Formats	Missing Values	Partial or Total Botnet Samples
			popularly called scenarios. The dataset is publicly available for download. It is multi class, has high class imbalance, mixed data types and real-life network traces						

The CTU-13 dataset covered in this study is summarised as shown in table 18. The pieces of information contained in the table are summary of the chosen dataset in this study.

4.2.3 Comments on Data Cleaning Methods for datasets used in ML-based Botnet Detection Models

To be able to use the dataset to build ML models, it is suggested that security researchers should make good efforts at identifying which of the shallow learning algorithms (single learners, ensembles, hybrid) or deep learning algorithms that can be most suitable for the botnet attack identification by using variety of standard approaches. The summary of the dataset characteristics provided in the table 5 will serve as actionable insights to researhers using the dataset to build better botnet detection models as addressing some of the issues will lead to novel and improve detection of botnet evidence. Depending on the patterns as well as some other identified characteristics in the dataset, performance results of models using metrics such as accuracy, precision, f1-score, recall e.t.c of the chosen algorithms may differ. The missing values in the dataset, the mixed data types, extreme class imbalance of the dataset, the multi-class nature of the attacks, the large samples in each of the captures will require that security researchers using the dataset for intrusion detection studies have to be very creative. Varying the split train-test ratio as well as changing the hyper parameters of the learning algorithms can equally help in building models that can adequately identify the botnet threats better.[16] stressed the need for researchers investigating botnet classification to be conscious of the detection evasiveness property of botnet. In each of the scenarios, it was observed that the number of missing values is very large and deleting them may introduce bias to the classification models as argued by [17] since some of the captures have hundreds of thousands of missing values. Thus, it is suggested that studies that want to use the CTU-13 dataset to build botnet detection models can handle the missing values in each of the thirteen scenarios through appropriate imputation method(s). Using this approach, the number of original samples in the dataset remains as the missing values have been filled with the imputed values. It is argued herein that the approach is better than deleting the missing values which may bring bias into the botnet classification results. It is also suggested that since the dataset contains mixed data (numerical and categorical), appropriate feature encoding technique should be employed by any researcher who intends using it.

4.2.4 Note on the Importance of EDA in Achieving Improved Attack Classification Models

EDA has been identified to be very promising in ML studies. The exploratory analyses carried out in this study has confirmed that the claims by Researchers in [18] as well as those in [19] are relevant. In fact, the claims further showed that all the data captures in CTU-13 dataset have

extreme class imbalance. Going by all the analyses in this study, it is evident that any researcher using CTU-13 for machine learning-based botnet detection will need to address some of the key issues in the dataset. The EDA step in the machine learning workflow is the one that reveal various characteristics of the dataset. Each of the captures in the dataset is huge and require that the researcher finds the best way to build less computationally expensive models from them. Based on some of the issues observed in the dataset, researchers will need to subject the dataset to various stages of pre-processing. Also, from the EDA, it was observed that the thirteen different captures of the dataset have various botnet types and instances.

On the strengths of the CTU-13 dataset for security researches, we agreed in this study with the argument of authors in [20] who defined a good dataset on security studies based on some established metrics in literature. For example, authors in [20] pointed out that a good Intrusion Detection dataset should have a minimal set of features, expected to be recent and should be able to realistically model the current internet traffic as it contains attack traces that are typical for modern networks and malware.

5. DISCUSSION OF THE FINDINGS

This study reported relevant information about botnet datasets in general and then CTU-13 dataset in particular. Exploratory data analyses of the thirteen captures in the CTU-13 dataset were carried out. The analyses procedure include: dataset description, computing the Statistical Summary and then identification of the basic properties. The statistical summary of all the thirteen scenarios are as shown in tables Generally, descriptive statistics is a way of providing brief overview of datasets including some measures and features of the samples. This is because it has been established that summary statistics are useful in data analytics. For instance, it is used for to spot patterns and then observe the general trends, like what the average is, how spread out the numbers are, and if there are any unusual numbers in the dataset.

The CTU-13 dataset is grouped in thirteen different captures, each with various thousands/millions of samples and a number of attacks that are found to be very destructive in networks. Looking at the number of packets and netflows in the dataset as shown in table 2, it can be deduced that to train a ML model for the classification of botnets in each of the scenarios (captures) requires some level of innovation so as to ensure that the models built are computationally less expensive.

The exploratory data analyses also exposed the hidden patterns as well as the relationships that exist among the features in the dataset. For instance, it was clearly discovered that the dataset suffers from extreme class imbalance, contains mixed data types as well as complex patterns that make its pre-processing very challenging. Though the feature space is not too large, the samples are in millions, making the dataset to be very large. Also, the dataset should also be sufficiently rich, capturing possibly all different kinds of malware behavior or attacks. Based on the results of the exploratory analyses in this study, it can be argued that the CTU-13 dataset is large, contains enough traffic information, is found to be representative and good for security studies. This is evident as shown in various summary statistics and dimensions obtained in tables 4 to 17 as part of the results of this study. Going by the different mentions of samples and attacks that are available in the CTU-13 dataset for botnet studies, it is evident that the dataset is good enough and can serve as representative for attack-related issues involving botnets.

6. CONCLUSION

This study reported an overview of CTU-13 botnet dataset. The study pointed out that the dataset is a publicly available one and is widely being used for botnet detection studies. This work equally pointed out that out of all the available botnet datasets, CTU-13 is currently the largest as it contains only real botnet and non-botnet. All the thirteen captures in the chosen CTU-13 dataset can be said to be representative for intrusion detection studies in the area of botnets. However, despite the promises that CTU-13 dataset has over some other ones, it is observed that it has its limitations which have to be addressed before being used to build ML models. In particular, exploratory investigation of the CTU-13 dataset revealed that it has the following drawbacks: it is very huge, contains mixed data types and its classes are highly imbalanced. High class imbalance

always reduce the ability of classification algorithms to adequately identify important cases and this can lead to mis-classification of positive samples as negative classes or the minority class may be treated as noise as argued in many literature. This is one of the reasons why this study recommends that the data imbalance problem is addressed prior to classification so that improved detection of botnet evidence can be arrived at. This study established that the CTU-13 dataset has mixed data types with complex data distributions and this is a justification on the need for adequate pre-processing of the dataset prior to carrying out classification.

On the positive side, detailed investigation of the features and samples in all the captures of CTU-13 dataset showed that the dataset is labeled and very promising for supervised learning-based botnet detection studies. Also, out of the most popular botnet datasets, CTU-13 is found larger, representative, contains real-life traces as well as more samples of different botnet malware as argued by authors who released it. Thus, it can be concluded from this study that CTU-13 dataset is a good dataset that can be found useful in various security studies that focus on the identification of the presence of botnet-based attacks in networks. The results of the exploratory analyses showed that all the thirteen different data captures in the CTU-13 dataset absolutely contains botnet samples only. Thus, it can be argued that the dataset is representative enough for ML-based botnet detection studies. It is concluded that the exploratory analyses can guide future researches in building improved botnet detection models using the CTU-13 dataset.

REFERENCES

[1] Z. Lei, Y. Shui, W. Di & Paul Watters, "A Survey on Latest Botnet Attack and Defense", International Joint Conference of IEEE Trustcom-11/IEEE ICESS-11/FCST-11, 2011.IV, 2011

[2] A. Pektas & T. Acarman, "Effective Feature Selection for Botnet Detection Based on Network Flow Analysis". *Inter,2017*

[3] A. Pektaş, & T. Acarman, " Botnet detection based on network flow summary and deep learning". *International Journal of Network Management*, *28*(6), 1–15. https://doi.org/10.1002/nem.2039, 2018

[4] S. Garcia, M. Grill, J. Stiborek & A. Zunino, "An empirical comparison of botnet detection method", *Computers and Security Journal, Elsevier, 45, 100-123. http://dx.doi.org/10.1016/j.cose.2014.05.011 detection approaches., 247–255.* doi:10.1109/cns.2014.6997492, 2014

[5] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, & C. Kruegel, "Disclosure: detecting botnet command and control servers through large-scale netflow analysis", in Proceedings of the 28th Annual Computer Security Applications Conference, ACM, 2012, 129–138,2012

[6] S. M. Kasongo & Y. Sun Y., "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset". Journal of Big Data 7, 105 (2020). https://doi.org/10.1186/s40537-020-00379-6-*2866-1_3*, 2020

[7] M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy & S. Othman, "A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System", Asian Journal of Research in Computer Science 7(4): 14-33,DOI: 10.9734/ajrcos/2021/v7i430185, 2021

[8] A. Alenazi, I. Traore, K. Ganame, & I. Woungang, "Holistic Model for HTTP Botnet Detection Based on DNS Traffic Analysis". In: Traore I., Woungang I., Awad A. (eds) *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham, 2017*

[10] L. H. Clemmensen & R. D. Kjærsgaard, "Data Representativity for Machine Learning and AI Systems", retrieved from https://www.semanticscholar.org/reader/ac9dd0a22c31c4e12c6c48559c4d06e567dbeac8 on 23rd December, 2023

[11] A. M. Oyelakin & R. G. Jimoh, **"**Tree-Based Learning Models for Botnet Malware Classification in Real Life Sub-Sample Dataset", Innovative Computing Review, published by the School of Systems and Technology (SST), University of Management and Technology (UMT), Lahore, Pakistan, 3(2), 1-13, Dec, 2023

[12] A. M. Oyelakin**,** A. O. Ameen, T. S. Ogundele, T. T. Salau-Ibrahim, U. T. Abdulrauf, H. I. Olufadi, I. K. Ajiboye, S. Muhammad-Thani, & I. A. Adeniji, "Overview and Exploratory Analyses of CICIDS 2017 Intrusion Detection Dataset", Journal of Systems Engineering and Information Technology (JOSEIT), 2(2), 45-52. https://doi.org/10.29207/joseit.v2i2.5411, 2023

[13] A. Mashkanova, "Exploratory Data Analysis toward Cloud Intrusion Detection", *A Master Thesis submitted to University of Victoria for the award of M.Sc. Computer Science, 2019*

[14] H. A. Mohammad & B. Najla , "A Detailed Analysis of New Intrusion Detection Dataset", Journal of Theoretical and Applied Information Technology,15th September 2019. Vol.97. No 17, 2019
[15] R. Panigrahi & S. Borah, "A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems", International Journal of Engineering & Technology 7(3):479-482, 2018
[16] A. M. Oyelakin, T. T. Salau-Ibrahim, B. S. Ogidan, R. D. Azeez & I. K. Ajiboye, "Peer-to-Peer Botnets: A Survey of Propagation, Detection and Detection Evasive Techniques", *Fulafia Journal of Science and Technology, a Tetfund-funded Journal of Federal University, Lafia, Nassarawa State, Nigeria, 5(3):13-18, 2019*
[17] M. Swamynathan, "Mastering Machine Learning with Python in Six steps, A Practical Implementation Guide to Predictive Data Analytics Using Python", *DOI:10.1007/978-1-4842,2017*
[18] S. Harun, T. H. Bhuiyan, S. Zhang, H. Medal & L. Bian, "Bot Classification for Real-Life Highly Class-Imbalanced Dataset", *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress,*565–572. , 2017
[19] A. M. Oyelakin & R. G. Jimoh, "Towards Building an Improved Botnet Detection Model in Highly Imbalance Botnet Dataset-A Methodological Framework", *Middle East Journal of Applied Science & Technology*, 3(1), January - March 2020, *available at http://mejast.com/towards-building-an-improved-botnet-detection-model-in-highly-imbalance-botnet-dataset-a-methodological-framework.html, 2020*

[20] M. Malowidzki, P. Berezinski & M. Mazur, " Network Intrusion Detection: Half a Kingdom for a Good Dataset", Conference: NATO STO- IST-139 Visual Analytics for Exploring, Analysing and Understanding Vast, Complex and Dynamic Data retrieved from https://pdfs.semanticscholar.org/b39e/0f1568d8668d00e4a8bfe1494b5a32a17e17.pdf?_ga=2.237473350.756880770.1576358584-422052986.1572640169, 2015

[21] H. A. Gameng, B. B. Gerardo & R. P. Medina, "Modified Adaptive Synthetic SMOTE to Improve Classification Performance in Imbalanced Datasets",*2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Kuala Lumpur, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICETAS48360.2019.9117287, 2019
[22] C. Beyan & R. Fisher, "Classifying imbalanced datasets using similarity based hierarchical decomposition", *Pattern recognition, 48(5), 1653-16728, 2015*