



Overview and Exploratory Analyses of CICIDS 2017 Intrusion Detection Dataset

Oyelakin A. M^{1*}, Ameen A.O², Ogundele T.S³, Salau-Ibrahim T⁴, Abdulrauf U.T⁵, Olufadi H.I⁶, Ajiboye I.K⁷,
Muhammad-Thani S⁸, Adeniji I. A⁹

¹³⁴⁵Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria

²⁶⁸⁹Department of Computer Science, University of Ilorin, Ilorin, Nigeria

⁷ Computer Science Unit, Abdulraheem College of Advanced Studies, Adenagar, Jordan

*amoyelakin@alhikmah.edu.ng

Abstract

Intrusion detection systems are used to detect attacks on a network. Machine learning (ML) approaches have been widely used to build such intrusion detection systems (IDSs) because they are more accurate when built from a very large and representative dataset. Recently, one of the benchmark datasets that are used to build ML-based intrusion detection models is the CICIDS2017 dataset. The data set is contained in eight groups and was collected from the Data Set & Repository of the Canadian Institute of Cyber Security. The data set is available in both PCAP and net flow formats. This study used the net flow records in the CIDIDS2017 dataset, as they were found to contain newer attacks, very large, and useful for traffic analysis. Exploratory data analysis (EDA) techniques were used to reveal various characteristics of the dataset. The general objective is to provide more insight into the nature, structure, and issues of the data set so as to identify the best ways to use it to achieve improved ML-based IDS models. Furthermore, some of the open problems that can arise from the use of the dataset in any machine learning-based intrusion detection systems are highlighted and possible solutions are briefly discussed. The EDA techniques used revealed important relationships between the input variables and the target class. The study concluded that the EDA can better influence the decision about future IDS research using the dataset. Thus, improved machine learning-based intrusion detection systems can be built from the data set once it is well understood and pre-processed.

Keywords: Intrusion Detection, Data Set Exploration, Machine Learning, Dataset Preprocessing

1. Introduction

An intrusion detection system (IDS) is a protection mechanism for detecting network attacks on a network. Machine learning (ML) approaches have become popular for building such intrusion detection systems (IDSs) due to the limitations of signature-based detection schemes. ML is a sub-field of Artificial Intelligence that allows algorithms to learn from data and its applications have been found promising across many domains [1]. ML-based IDSs are more accurate when built from a very large and representative data set. Several machine learning-based intrusion detection systems have been proposed in the literature. These machine learning-based models have been built from different datasets. Some of such datasets include KDD CUP-99, NSL-KDD, Kyoto 2006+.

However, some of these data sets are old and have been used extensively in intrusion detection studies, while some are very small. In recent times, one of the benchmark data sets that are becoming popular to build ML-based intrusion detection models is the CICIDS2017 data set. The data set consists of eight different captures. Malowidzki Marek, Berezinski Przemyslaw, and Mazur Micha (2015) pointed out that a very large dataset that has representative attacks is better used for building intrusion detection models. The limitations observed in some of these datasets led to the design of the CICIDS2017 datasets as argued by Sharafaldin et al. (2018). Aside this, several works in the past have used these datasets to build machine learning-based intrusion detection systems.

This study specifically improves on a recent study that focused on exploratory analysis of some selected intrusion detection datasets. The work was authored by Ghurab, Gaphari, Alshami, Alshamy, and Othman (2021). However, the study did not detail the characteristics of the CICIDS2017 dataset. The approach used in this study is to perform a more detailed analysis of the CICIDS2017 dataset and then point out some of the open problems that researchers may face when using the dataset to build machine learning-based intrusion detection models. It is believed that this approach will be more comprehensive and can provide leading insights to researchers working in this area.

This paper focuses on reporting an overview of the data set and providing results of its exploratory analyzes. Komoroski, Marshall, and Saiccioli (2016) and Gibson and Freisas (2015) have argued that exploratory

analysis is a crucial step in every data analytics research, and this serves as the basis for the approach in this work. In any study based on machine learning, it is essential to identify the patterns that could be present in the chosen data set to know the best approach to using the data set for model building. Therefore, the focus of this study is to perform an overview and exploratory analysis (EDA) of the IDS data set. Generally, an EDA is the process of getting to know data in depth so as to have a better understanding of how to use it in building ML-based models.

Furthermore, exploratory data analysis enables machine learning researchers to remove irregularities, outliers, and unnecessary values from the dataset, thereby promoting the building of improved models in different domains. This paper first provided an overview of the different captures in the dataset and emphasizes the need to address the many features contained in each dataset capture. Then, this study used different EDA approaches to provide better insight into the data set and then discussed some of the challenges of using the data set in IDS studies. The general objective is to provide more information on the data set that can aid in the construction of improved ML-based IDS models.

2. Related Studies

Ghurab et al. (2021) performed an analysis of some benchmark data sets that are used to build network intrusion detection systems. The study generally discussed old and new datasets for IDS studies. However, it was observed that the analyzes were general and a detailed report was not made on a recent dataset named CICIDS2017. Similarly, Panigrahi and Borah (2018) carried out an analysis of the CICIDS2017 data set that is being recently used to build intrusion detection systems. The paper explored general characteristics of the data set and mentioned some of the inherent issues with respect to it without focusing on exploratory analyses. Aggarwala Preeti & Kumar Sharmab Sudhir (2015) carried out an analysis of the KDD CUP 99 dataset attributes class-wise for intrusion detection. The experimental analysis in the study revealed better insights on the KDD CUP dataset, which is also popular for intrusion detection studies. Apart from this, Proti (2018) conducted a review of three datasets, namely the KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets, which are popular for research on intrusion detection studies [2].

Iman, Arash, and Ali (2018) argued that some of the major limitations observed in the previous IDS dataset brought about the need for the development of the CICIDS2017 dataset. The authors carried out an analysis of the CICIDS2017 data set. The study discussed some of the key features and components of the dataset. However, the study did not reveal some issues from the analysis and did not extend to reporting the open problems found in the data set. Specifically, the authors claimed that their evaluations of about 11 previous datasets showed that most of them are out of date and unreliable. Some of them also suffer from the lack of diversity and traffic volumes, as they do not cover the variety of known attacks. Similarly, Mashkanova (2019) carried out Exploratory Data Analysis of Cloud-based Data Set that can be used for identifying intrusions in Cloud computing environment. The focus of the work was only on cloud computing security issues.

Komorowski et al. (2016) listed some tools used to explore a dataset, which is essential to gain a good understanding of the features and potential issues of the dataset. Gibson and Freitas (2015) presented the research contexts, the tools and methods used in the exploratory phases of the analysis, the main findings, and the implications for learning analytics research methods. Santosh, Sahu, Sarangi and Jena (2014) carried out an analysis of some intrusion detection datasets such as KDD-99, NSL-KDD, etc. The data sets used in the investigation were the ones that have been reported to be very old. Tavallae et al. (2009) conducted a statistical analysis on the KDD CUP 99 dataset and reported that there are two important issues that highly affect the performance of intrusion detection systems built with it. Therefore, the authors proposed a new data set named NSL-KDD, which consists of selected records of the entire KDD 99 data set but improved on the mentioned shortcomings of the old data set.

3. Methods

The data set used in this study was collected from the Canadian Institute of Cyber Security Data Sets repository. It is available for download at <https://www.unb.ca/cic/datasets/ids-2017.html>. The methods used in this study are two-fold. First, an overview of the intrusion detection dataset named CICIDS2017 was provided. Thereafter, the focus is on performing detailed exploratory analyzes of the eight different captures in the dataset. The data set was chosen because it is very large and contains several attacks and intrusion traces, which is good for security studies. The exploratory analyze procedure includes the following: dataset description, computing the statistical summary, identification of the properties in the datasets, and data visualization. Then some of the open problems of the data set identified in the exploratory analyses are discussed. All experiments were carried out in the Python programming language environment.

3.1 Data sets for Intrusion Detection Studies

Several data sets have been released for intrusion detection studies. In fact, they are too numerous to mention. Some of these datasets are listed below. They include: KDD CUP 99, NSL-KDD, IoT Healthcare Security Datasets, IoT DOS datasets, IoT DDOS Security datasets, Kyoto 2006+, datasets on malware of different types, and many others. In this work, CICIDS2017 is studied, which is one of the most popular IDS datasets in recent times, with a view to revealing some of the issues with it and how to use it to build IDS models better.

4. Results and Discussion

The findings of this study are grouped into two. The first reported an overview of the eight captures in the data set. The second results are based on detailed exploratory analyzes. Some of the open problems identified in the data set based on the EDA are also discussed.

4.1 Overview of the CICIDS2017 data set

From the analysis carried out, it was discovered that CICIDS2017 is a large and representative data set that is good for evaluating intrusion detection systems. The data set was originally developed at the Faculty of Computer Science; University of New Brunswick. The data set was built and released by Sharafaldin et al. (2018) purposely to advance studies on the building of intrusion detection systems. The data set contains up-to-date benign common attacks, which resembles the true real-world data (PCAP). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source and destination IPs, source and destination ports, protocols, and attacks.

The CICIDS2017 dataset consists of labeled network flows, including full packet payloads in pcap format, the corresponding profiles and the labeled flows that are publicly available for researchers (Sharafaldin et al., 2018). As argued by Sharafaldin et al. (2018), they built the abstract behavior of 25 users based on the HTTP, HTTPS, FTP, SSH, and email protocols in the dataset. The authors pointed out that the data capture period for the CISIDS2017 data set started at 9 am on Monday, 3 July 2017 and ended at 5 p.m. on Friday, 7 July 2017, for a total of 5 days. Also in the dataset, the available attacks include Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS.

During the data set building, the attacks were executed both morning and afternoon on Tuesday, Wednesday, Thursday, and Friday. There are eight different captures in the data set. Each of these captures contains attacks recorded during the data set building. Based on the period of capture in those periods, the different captures in the data set were renamed in this study FriAfternoonPortScan, FriAfternoonDDOS, FriMorning, MonMorningHour, ThurAfternoonInfiltration, ThursdayMornWebAttacks, TueWorking, WedHour for easy referencing purposes.

4.2. Results of Exploratory Data Analysis

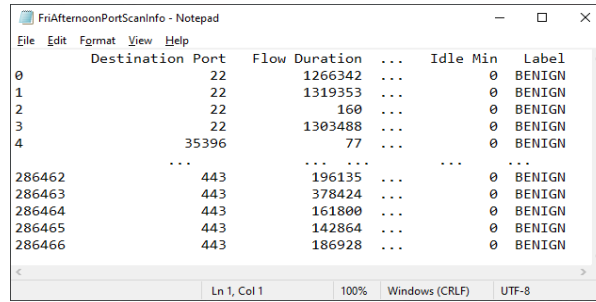
Table 1. Dataset feature space and sample size

Capture	Name Chosen for The Data Set Capture	No Input Features	No Samples/Instances
Capture 1	FriAfternoonPortScan	78	286,467
Capture 2	FriAfternoonDDOS	78	225,745
Capture 3	FriMorning	78	191,033
Capture 4	MonMorningHour	78	529,918
Capture 5	ThurAfternoonInfiltration	78	288,602
Capture 6	ThursdayMornWebAttacks	78	170,366
Capture 7	TueWorking	78	445,909
Capture 8	WedHour	78	692,703

The experimental results obtained in Table 1 are the true description of the features and instances in the CICIDS2017 data set.

4.3. Data Distributions in the Dataset Captures

The data distributions in the data set are as shown in Figure 1:

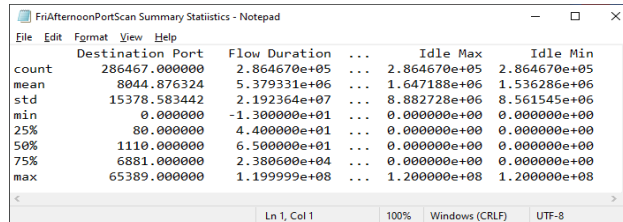


	Destination Port	Flow Duration	Idle Min	Label
0	22	1266342	0	BENIGN
1	22	1319353	0	BENIGN
2	22	160	0	BENIGN
3	22	1303488	0	BENIGN
4	35396	77	0	BENIGN
286462	443	196135	0	BENIGN
286463	443	378424	0	BENIGN
286464	443	161800	0	BENIGN
286465	443	142864	0	BENIGN
286466	443	186928	0	BENIGN

Figure 1. Data frame of the first data capture

4.4 Summary statistics in the dataset captures

The statistical summary provides some statistical details about the distributions in the chosen dataset. The summary statistics of the eight captures in the data set are shown in Figure 2.



	Destination Port	Flow Duration	Idle Max	Idle Min
count	286467.000000	2.864670e+05	2.864670e+05	2.864670e+05
mean	8044.876324	5.379331e+06	1.647188e+06	1.536286e+06
std	15378.583442	2.192364e+07	8.882728e+06	8.561545e+06
min	0.000000	-1.300000e+01	0.000000e+00	0.000000e+00
25%	80.000000	4.400000e+01	0.000000e+00	0.000000e+00
50%	1110.000000	6.500000e+01	0.000000e+00	0.000000e+00
75%	6881.000000	2.380600e+04	0.000000e+00	0.000000e+00
max	65389.000000	1.199999e+08	1.200000e+08	1.200000e+08

Figure 2. Summary statistics for the first data capture

From the summary statistics, it was observed that the distributions are similar on the basis of the values obtained from each statistical result. This further confirms that the intrusion captures in each of the net flow dataset behaved in a similar manner.

4.5 Visualization of patterns in the data set

The visualization of each of the sets in the dataset is captured as shown in Figures 17 to 24. They are all basic scatter plots that represent the patterns in the dataset, the more.

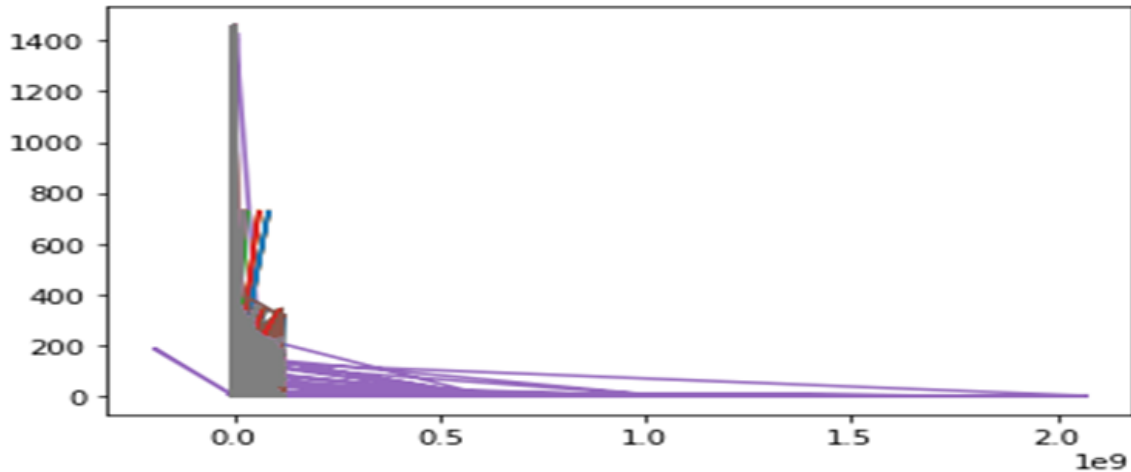


Figure 3: Visualization of First Data Capture

Statistical summaries and diagrams are used to show the description of the patterns in the dataset. For example, it can be seen from Figure 1 that different patterns exist from the eight sets of data set capture. The

spread across the X and Y axes differs. Generally, scatter plots help ML researchers identify correlations between variables and estimate the nature of the correlations.

Table 2. Key Summary of the Captures in the CICIDS2017 Dataset

S/N	Dataset Name	Brief Information about the Dataset	Class Distribution	Data Types	Attack Types	Available formats	Unknown or Missing Values?	Open Problems Observed in the Data Set
1	8Captures CICIDS2017 data set	The data set is usually used as a reference for intrusion detection studies.	Imbalance class distribution	Numeric (a mixture of integer and floating-point data types)	Several types of attacks of different magnitudes are captured in the data set.	NetFlow and PCAPs. The format used for all of the experimental analyses in this paper is net flow.	YES, there are few unknown or missing values in each group of the dataset. That is, each of the sets of datasets has Nan values that have to be addressed before being used to build models.	The data set is very large and contains complex data patterns and high-class imbalance, the input variable values are on different scaling.

Generally, the summary of the features and samples in the eight captures of the data set is summarized as shown in Table 2.

Table 3. Summary of Suggested Solutions for tackling the Issues in the Dataset

Dataset Name	Brief Information About the Dataset	Solving Unbalanced Class Distribution	Scaling The Features Due to Data Types	Choosing The Right Dataset Formats.	Handling Unknown or Missing Values	Suggested Solutions to The Open Problems Observed in The Data Set
CICIDS2017 data set	The data set is a reference for intrusion detection studies.	Imbalance class distribution. Using the dataset in its raw form without addressing the class imbalance will make the ML based models built from it biased. The minority class may need to be improved using the Synthetic Minority Over Sampling Technique (SMOTE) proposed by Chawla et al. (2002) and techniques proposed by Beya & Fisher (2015), as well as Gameng, Gerardo, and Medina (2019) can be of great help. Any other related techniques as found in the literature that are suitable can also be used to address the challenge of high-class imbalance.	Since the dataset contains input features with a mixture of integer and floating-point data types, each with different range of values, there may be a need to do scaling or normalization. This may be necessary as a result of some types of machine learning algorithms that may not perform effectively if such data are fed into them.	Researchers may consider using the net flow data because it contains no packet header and may be better than the PCAPs formats. In this study, the format used for the experimental analysis in this paper is net flow.	The missing values can be handled by deletion or imputation. All depends on the technique for which the researcher intends to settle, and this must be justified. The main disadvantage of skipping or deleting missing values is that important information needed by the machine learning model may be deleted. Therefore, the results will be biased. Therefore, the argument of deleting missing values if the data set is very large and missing values are not more than 5% cannot be supported in some circumstances.	For this reason, handling the big data issue while using the data set is required to build intrusion detection models is required. Furthermore, it was observed that the data set has complex data patterns. Thus, machine learning algorithms that have the ability to handle complex distributions have to be chosen when building machine learning-based models. Another open problem in the data set that has to be addressed is the high-class imbalance. The other issue is that the data set has many features that cannot be used to build the model. Therefore, as argued by Oyelakin and Jimoh (2021), the selection of features will be very essential. This approach will allow researchers to build an ML-based intrusion detection model based on the reduced features in the CICIDS2017 dataset. Thus, the models will be less complex, more interpretable, and will have excellent performance. Lastly, proper scaling of the dataset features has to be addressed, as well, because of the high variation in some of the feature scaling.

Table 4: Records of each data set before and after that were deleted when unknown or missing values were deleted

Capture	Name Chosen for The Data Set Capture	No Input Features.	No Original Samples/Instances	No Unknown or Missing Values (Data)	Comment On the Deleted Values
Capture 1	FriAfternoonPortScan	78	286,467	015	The missing value in this capture of the data set is very minimal.
Capture 2	FriAfternoonDDOS	78	225,745	004	The missing value in this capture of the data set is very minimal.

Capture 3	FriMorning	78	191,033	028	The missing value in this capture of the data set is very minimal.
Capture 4	MonMorningHour	78	529,918	064	The missing value in this capture of the dataset is fairly large.
Capture 5	ThurAfternoonInfiltration	78	288,602	018	The missing value in this capture of the data set is very minimal.
Capture 6	ThursdayMornWebAttacks	78	170,366	020	The missing value in this capture of the data set is very minimal.
Capture 7	TueWorking	78	445,909	201	The missing values in this capture of the data set are very large. Of course, it is the largest among the eight captures.
Capture 8	WedHour	78	692,703	008	The missing value in this capture of the data set is very minimal.

This study used a recent and rich intrusion detection data set named the CICIDS2017 dataset for experimental analyzes. First, an overview of the data set was reported. The focus was then shifted to the use of different exploratory data analysis (EDA) approaches to get a better understanding of the data set. The study first revealed the different data frames in the data set. Subsequently, summary statistics were obtained for each set of data set captures. The statistical summary provided essential statistical information about the characteristics and samples of the data set. From the EDA, it was also discovered that there are 79 missing (NaN) values in each of the dataset captures.

Aside this, analyses revealed that the input features (attributes) in the dataset are of numeric data type (integer and floating types) while the output feature is categorical (Benign and non-benign). On the basis of the exploratory analysis of the dataset, it was equally found that the input features are of different values and ranges. The data set was also observed to have a high-class imbalance. This study observed that the features in the dataset have complex data patterns, which require innovative approaches during the pre-processing stages so as to be able to build more effective intrusion detection models from the dataset. It was equally discovered that the eight different captures in the data set reported various attacks, and the numerical data are of integer and floating-point type. The exploration revealed the structure of the dataset, some of the problems that need to be addressed, and better approaches to address the dataset shortcomings in a machine learning classification problem.

Some of the issues identified with the data set are summarized in Table 2. For example, since some of the ML-based IDS cannot learn from a data set with missing values, the issue has to be addressed. The popular arguments for handling missing values include: deleting the columns whenever missing values are found, using imputation (mean or mode imputation). For instance, Swamynathan (2017) pointed out that once a data set is very large and the missing values are less than 5%, the missing ones can be deleted. This study agrees with this argument, since the CICIDS2017 dataset is very large, running to several gigabytes of information and the unknown (missing) values are very minimal.

This study hereby recommends that researchers using the dataset may consider imputation techniques to handle the unknown or missing values and then use the preprocessed dataset to build an improved ML-based intrusion detection system. Further analysis carried out showed that there is a need to address the class imbalance in each of the capture using any suitable method in the literature. Some of the summarized solutions are mentioned in Table 3 and can be of great help to any machine learning researcher who proposes to use the data set for IDS studies. Table 4 was used to present the results of the experimental analysis of the data set with respect to the feature space and sample sizes before and after the removal of missing values. Visualizations of the data set carried out in the study also provided some insight into pattern distributions. It is believed that understanding the distributions can help researchers better use the data set in future research.

5. Conclusions

This study used innovative approaches to provide a detailed analysis of the data set. The work focused on investigating the basic characteristics of the benchmark intrusion detection data set named CICIDS2017 using some exploratory data analysis techniques. The data set used in this study was collected from a repository in a Canadian university laboratory. The experimental analyzes of the data set are detailed and can provide adequate information to researchers using it to build intrusion detection systems. The patterns in the data set were also visualized using a simple scatter plot. It is believed that the exploratory data analysis further revealed some of the underlying structures/patterns in the data set, which can help build improved ML-based intrusion detection models. Equally important, some of the suggestions made in this study to handle open problems in the data set can serve as information for researchers working in the IDS area. The EDA techniques used in this study may be useful to reveal important relationships between input variables and the target class. The study concluded that the EDA can better influence the decision about future IDS research using the dataset. A future study will focus on building efficient ML-based models from the CICIDS2017 dataset, with an emphasis on the impact of innovative data cleaning approaches on the performance of the targeted ML models.

References

- [1] Aggarwala Preeti & Kumar Sharmab Sudhir (2015). Analysis of KDD Dataset Attributes - Class-wise For Intrusion Detection, 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015), Procedia Computer Science 57, 842 – 851
- [2] Beyan C. & Fisher, R. (2015). Classifying imbalanced datasets using similarity-based hierarchical decomposition, Pattern recognition, 48(5), 1653-16728
- [3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. <https://doi.org/10.1613/jair.953>
- [4] Gameng H.A., Gerardo B. B. & Medina R. P., (2019). 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, 1-5, doi: 10.1109/ICETAS48360.2019.9117287.
- [5] Ghurab Mossa , Gaphari Ghaleb, Alshami Faisal, Alshamy Reem & Othman Suad (2021). 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
- [6] Gibson David C & Freitas Sara de (2015). Exploratory Analysis in Learning Analytics, Technology, Knowledge, and Learning 21(1), DOI: 10.1007/s10758-015-9249-5
- [7] Komorowski Matthieu Marshall Dominic C. , Salciccioli Justin D & Crutain Yves (2016). Exploratory Data Analysis, In book: Secondary Analysis of Electronic Health Records, 10.1007/978-3-319-43742-2_15
- [8] Malowidzki Marek, Berezinski Przemyslaw & Mazur Micha (2015). 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
- [9] Mashkanova Aigerim (2019). Exploratory Data Analysis toward Cloud Intrusion Detection, A Master Thesis submitted to University of Victoria for the award of M.Sc. Computer Science
- [10] Mohammad Hamid Abduraheem & Najla Badie Ibraheem (2019). A Detailed Analysis of New Intrusion Detection Dataset, Journal of Theoretical and Applied Information Technology 15th September 2019. 97(17)
- [11] Oyelakin A.M. & Jimoh R.G. (2021), A Survey of Feature Extraction and Feature Selection Techniques Used in Machine Learning-Based Botnet Detection Schemes, VAWKUM Transactions on Computer Sciences, 9 (2021),1-7, available at <https://vfast.org/journals/index.php/VTCS/article/view/604/658>
- [12] Panigrahi Ranjit & Borah Samarjeet (2018). A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems, International Journal of Engineering & Technology, 7(3):479-482
- [13] Protić Danijela D.(2018). Review of Kdd Cup '99, Nsl-Kdd and Kyoto 2006+ Datasets, Military Technical Courier, 66(3), DOI: 10.5937/vojtehg66-16670; <https://doi.org/10.5937/vojtehg66-16670>
- [14] Sharafaldin I., Lashkari A. H. Ghorbani A. A. (2019). A Detailed Analysis of the CICIDS2017 Data Set. Springer, International Conference on Information Systems Security and Privacy, 2019.
- [15] Smola Alex & Vishwanathan S.V.N. (2008). Introduction to Machine Learning Cambridge university press. The Edinburgh Building, Cambridge, UK
- [16] Santosh Kumar Sahul, Sauravranjan Sarangi & Sanjaya Kumar Jena(2014). A Detail Analysis on Intrusion Detection Datasets, 2014 IEEE International Advance Computing Conference (IACC)
- [17] Sharafaldin Iman, Lashkari Arash Habibi, and Ghorbani Ali A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018
- [18] Swamynathan Manoha (2017). Mastering Machine Learning with Python in Six steps, A Practical Implementation Guide to Predictive Data Analytics Using Python, DOI:10.1007/978-1-4842-2866-1_3, or <https://tanthiamhuat.files.wordpress.com/2018/04/mastering-machine-learning-with-python-in-six-steps.pdf>

- [19] Tavallace M, Bagheri E., Lu W. & Ghorbani A. A. (2009). A Detailed Analysis of the KDD CUP 99 Dataset, Proceedings of the 2009IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)