Data Sets for Cyber Security Machine Learning Models: A Methodological Approach

Innocent Mbona[a] and Jan H. P. Eloff[b]
Department of Computer Science, University of Pretoria, Lynwood Road, Pretoria, South Africa

Keywords: Cyber Security, Real-World Data, Synthetic Data, Zero-Day Attack, Network Intrusion Detection Systems.

Abstract: Discovering Cyber security threats is becoming increasingly complex, if not impossible! Recent advances in artificial intelligence (AI) can be leveraged for the intelligent discovery of Cyber security threats. AI and machine learning (ML) models depend on the availability of relevant data. ML based Cyber security solutions should be trained and tested on real-world attack data so that solutions produce trusted results. The problem is that most organisations do not have access to useable, relevant, and reliable real-world data. This problem is exacerbated when training ML models used to discover novel attacks, such as zero-day attacks. Furthermore, the availability of Cyber security data sets is negatively affected by privacy laws and regulations. The solution proposed in this paper is a methodological approach that guides organisations in developing Cyber security ML solutions, called CySecML. CySecML provides guidance for obtaining or generating synthetic data, checking data quality, and identifying features that optimise ML models. Network Intrusion Detection Systems (NIDS) were employed to illustrate the convergence of Cyber security and AI concepts.

1 INTRODUCTION

The convergence of artificial intelligence (AI) and Cyber security is high in the research agenda of AI and Cyber security experts. In general, the state of Cyber security could benefit from advances made in AI. Consider for example the discovery of insider threats, such as disgruntled employees, who may damage the reputation of an organisation. Insider threats are difficult, if not impossible to discover using techniques such as monitoring and auditing. Alternatively, if you have a well-trained and tested machine learning (ML) platform, the intelligent discovery of insider threats becomes a reality! However, progress made in the successful convergence of Cyber security and AI is hampered by the unavailability of relevant and sufficient volumes of real-world attack data sets. Furthermore, most organisations do not have access to useable, relevant, and reliable real-world attack data, such as network traffic data, to unlock the benefits of AI. In addition, the availability of quality data sets that can be efficiently used to train ML algorithms is negatively impacted by privacy laws and regulations (Lu et al., 2023).

This shortage of data sets that can be used on Cyber security ML platforms is now demanding that researchers, Cyber security professionals, and organisations search for alternative data sets. The Financial Times (Financial Times, 2023) reported that AI companies are considering innovative ways to address the shortage of data sets. Furthermore, it was suggested that companies should consider generating useable, relevant, and reliable data themselves. However, it should be noted that generating your own data for ML environments are likely to increase the risk exposure of an organisation in the sense that the data can be biased towards some goal, and therefore miss real-world attack scenarios (Figueira & Vaz, 2022; Lu et al., 2023).

Existing literature lack a structured approach that can guide organisations and Cyber security professionals to ensure that important aspects such as using quality data and identifying significant features are adequately implemented, particularly for discovering complex scenarios such as zero-day attacks. This paper optimises existing techniques that

[a] https://orcid.org/0000-0001-9240-8035
[b] https://orcid.org/0000-0003-4683-2198

Mbona, I. and Eloff, J.
Data Sets for Cyber Security Machine Learning Models: A Methodological Approach.
DOI: 10.5220/0012598400003705

In *Proceedings of the 9th International Conference on Internet of Things, Big Data and Security (IoTBDS 2024)*, pages 149-156
ISBN: 978-989-758-699-6; ISSN: 2184-4976

include high data quality checks, feature selection and ML models to address this problem.

The remainder of this paper address the following research questions:

(i) Given the scarcity of real-world attack data, how to create or obtain Cyber security data sets for ML?

(ii) What data quality aspects should be considered for Cyber security data sets used in ML?

(iii) What are the important features to monitor on Cyber security data sets to effectively discover complex attacks such as zero-day attacks?

2 LITERATURE REVIEW

The literature overview and related work section is structured as follows:

- Obtaining Cyber security data sets for ML.
- Ensuring the quality of Cyber security data sets.
- Performing feature selection on Cyber security data sets.

2.1 Data Collection: Cyber Security Data for Machine Learning

Obtaining Cyber security data for ML refers to the methods used for collecting data to be used in the experiments of a ML model (Abdallah & Webb, 2017; Kilincer et al., 2021). ML models rely on input data used in the training and testing stages when designing a ML model (Bonaccorso, 2020; Van Der Walt & Eloff, 2018). It is important for a ML model be trained on a data set that is representative of the problem at hand (Amr, 2019; Bonaccorso, 2020), so that once deployed, it can produce reliable results. Specifically, if a ML model is trained on a flawed data set, then a ML model is prone to producing unreliable results (Brownlee, 2020).

ML based Cyber security solutions should be trained on real-world attack data (Kilincer et al., 2021). However, there are multiple shortcomings regarding obtaining real-world data for ML based Cyber security solutions: (i) lack of large real-world attack data, (ii) sensitive nature of the data, and (iii) security, confidentiality and privacy concerns (Bowles et al., 2020; Lu et al., 2023). According to Singh et al. (Singh et al., 2019) privacy and security concerns are the main factors that cause organisations to not publicly share their real-world attack data. Anonymisation, which is a method of removing identity related or sensitive information from a data set, is a technique that can be considered to overcome privacy and security challenge (Dankar & Ibrahim, 2021); however, it does not address the data size challenge for the purposes of training a ML model (Bowles et al., 2020; Kumar & Sinha, 2023).

To overcome these challenges, researchers have proposed using synthetic attacks (Figueira & Vaz, 2022; Lu et al., 2023). Synthetic attacks are generated in a controlled environment such as a laboratory, usually based on a sample of real-world attacks using data generating tools (Bowles et al., 2020; Dankar & Ibrahim, 2021). An example of such a tool is CICFlowMeter (Kaushik & Dave, 2021; Ullah & Mahmoud, 2020) that generates network traffic flows.

Synthetic data should depict the true characteristics of real-world data and be privacy-preserving (Dankar & Ibrahim, 2021). Specifically for this paper, the question is, how to obtain or create Cyber security data representative of novel zero-day attacks? The problem is that existing literature lack a standardised approach of training and testing ML Cyber security solutions for discovering zero-day attacks, while ensuring that the data sets used are of high-quality. For instance, Zoppi et al. (Zoppi et al., 2021) and Pu et al. (Pu et al., 2021) proposed creating zero-day attack scenarios by using a particular set of known (synthetic) attacks in the training phase and use a different set of known attacks in the testing phase of a ML model. Their argument is that an attack that appears only in the testing phase, but not in the training phase of a ML model can be treated as a zero-day attack (Pu et al., 2021; Zoppi et al., 2021). However, with this approach it is not clear how best to choose attacks to use in the training and testing stages to optimise a ML model. This problem of obtaining or creating data for zero-day attacks is addressed in Section 3.1.

2.2 Assess the Data: Checking Cyber Security Data Quality

There are vast descriptions of data quality in the literature across different fields of study and industries. The characteristics to be considered are mainly informed by the requirements of an application system that takes the data as input. According to Mahanti (Rupa Mahanti, 2019) the most common characteristics of data quality include accuracy, reliability, legitimacy, consistency, relevance, flexibility, currency, completeness, availability, uniqueness, and timeliness amongst others. These characteristics of data quality are not explicitly

defined for a ML or AI problem. However, characteristics such as accuracy and relevance are of prime importance for ML problem solving in particular for Cyber security solutions. CySecML methodology aims to, amongst other things, provide guidelines of data quality checks to be considered for ML based Cyber security solutions. Cai et al. (Cai & Zhu, 2015) presented a framework of data quality characteristics that consists of: availability, usability, reliability, relevance, and presentation quality.

Although all the data quality frameworks mentioned in the previous few paragraphs are equally applicable to the different application domains, the research at hand adopted the Cai et al. (Cai & Zhu, 2015) framework. The main reason for using Cai et al. (Cai & Zhu, 2015) is for simplicity and Cyber security application domain needs.

2.3 Feature Selection: Feature Selection on Cyber Security Data

Various attributes and features describe activities in various application domains, such as NIDS (Aljawarneh et al., 2018; Kurniabudi et al., 2020; Moustafa & Slay, 2017). The information contained in attributes and features can be used for analysis, including input variables into ML models (García et al., 2015; Khalid et al., 2014; Mukkamala & Sung, 2003). However, not all features are useful or significant in the design of ML models.

Given the large number of attributes and features to be potentially found in Cyber security data sets, such as NIDS, the question is which features should be monitored to discover malicious activities, in particular complex malicious attacks such as zero-day attacks? A feature or attribute is considered significant if it can differentiate two or more data classes (Mukkamala & Sung, 2003).

A plethora of feature selection methods exists, see (Mbona & Eloff, 2022b; Zheng & Casari, 2018). For the research at hand Benford's law (BL) is presented as a feature selection method to assist ML models used in discovering Cyber security threats, see (Mbona & Eloff, 2022a, 2022b) for further details. BL is proposed as a feature selection method within the CySecML methodology given its unique properties for identifying anomalous behaviours in Cyber security data sets (Mbona & Eloff, 2022a, 2022b).

3 CySecML METHODOLOGY

The overall aim of the CySecML methodology is to assist organisations and researchers to follow a step wise process for developing ML based solutions for Cyber security related problems. For the purposes of this paper, the CySecML methodology consists of two components: data preparation and InternetBotDetector (IBD). The data preparation component consists of data collection or data preparation, data quality checks, data cleaning, feature selection and significant features. The IBD model is based on ML models that assist in the automation of discovering cyber attacks.

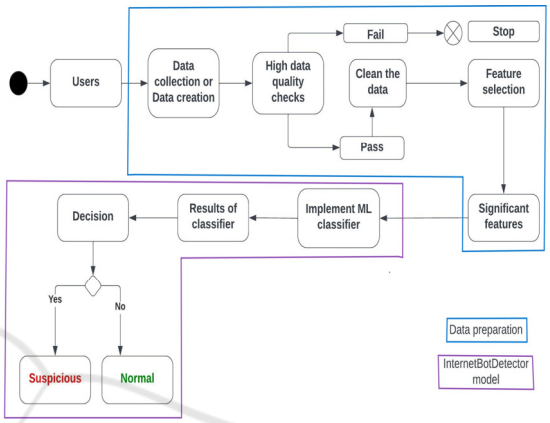

Figure1: Overview of the CySecML methodology.

The Cyber security data set that is used for illustrative purposes to highlight and explain the application of the CySecML methodology is the IoT intrusion-2020 (Ullah & Mahmoud, 2020) data set. A brief description of the IoT intrusion-2020 data set follows:

Ullah et al. (Ullah & Mahmoud, 2020) presented a network intrusion data set, called IoT Intrusion-2020, based on synthetic attacks. This data set was developed in 2020 based on various synthetic botnet attacks that include flooding attacks amongst others, and benign network traffic. In addition, the authors of this data set used the CICFlowMeter tool (Kaushik & Dave, 2021; Ullah & Mahmoud, 2020) to generate network traffic data.

3.1 Data Creation: Cyber Security Data for Zero-Day Attacks

The CySecML methodology includes a specific step for the process of Cyber security data collection or data creation (see Figure 1). To illustrate concepts relating to Cyber security data collection the discussion that follows focuses on collecting network traffic and network intrusion data. Cyber security data collection refers to obtaining relevant data for running ML models.

It is well known that real-world Cyber security data sets are most of the time not available due to confidentiality and privacy reasons. Most organisations and researchers are using synthetic data that is based on real-world data. For the purposes of this paper and in particular to develop ML models that are able to discover zero-day network intrusion attacks, the following publicly available Cyber security data sets were considered: the IoT intrusion-2020 (Ullah & Mahmoud, 2020), UNSW-NB15 (Moustafa & Slay, 2017), CICDDOS2019 (Sharafaldin et al., 2019), and CIRACICDOHBRW2020 (Montazerishatoori et al., 2020). In general, and particularly referring to the network intrusion data sets, the attacks considered in these data sets were synthetically generated. However, they do not contain zero-day (i.e., unknown) attacks. Specifically, for the purposes of this paper, a known attack is an attack that is found in a particular data set with its label. Whereas an unknown attack is a type of attack not part of a particular data set (Mbona & Eloff, 2022a).

Now the question is, how do we collect, create, obtain or fabricate data that is representative of unknown attacks? The rule of thumb in creating a zero-day attack type from known attacks is to not use all known type of attacks in the training and testing stages of a ML model (Ahmad et al., 2023; Zavrak & Iskefiyeli, 2020; Zoppi et al., 2021). Using all available network intrusion type of attacks in both the training and testing phase of a ML model disregards the presence of zero-day attacks (Ahmad et al., 2023), thus this approach is not preferred. According to Ahmad et al. (Ahmad et al., 2023), the two most common approaches for creating a zero-day attack type from known attacks are:

• Approach 1: randomly use different type of attacks in the training and testing stages of a ML model.

• Approach 2: combine all known attacks from a data set and create a new zero-day attack type that is based on known type of attacks.

Approach 1

The first approach for creating a zero-day type of attack for experimental purposes is to train and test a ML model using different type of attacks that are known (Zavrak & Iskefiyeli, 2020; Zoppi et al., 2021). As far as the authors of this paper are concerned, there is no scientific evidence in the literature on choosing the "optimal" combination of type of attacks to use as a zero-day attack type. In addition, consider the IoT Intrusion-2020 data set for example that contains eight malicious network traffic types, one can have a minimum of: 8!/(8-2)! = 28, possible combinations of zero-day attack types. Therefore, the authors of this research do not consider this approach feasible.

Approach 2

The second approach is to create a new unknown attack type (i.e., zero-day) by combining known attack types of a particular data set (Ahmad et al., 2023; Mbona & Eloff, 2022a). In this way, a zero-day attack type is created from known type of attacks and this approach is referred to as creating *known unknowns* (Ahmad et al., 2023; Kim et al., 2020; Scheirer et al., 2014). The premise of this approach is that a network traffic data set is based on the same description of features and structure, thus, one can combine different known type of attacks across the same features. "Approach 2" is depicted in Figure 2.

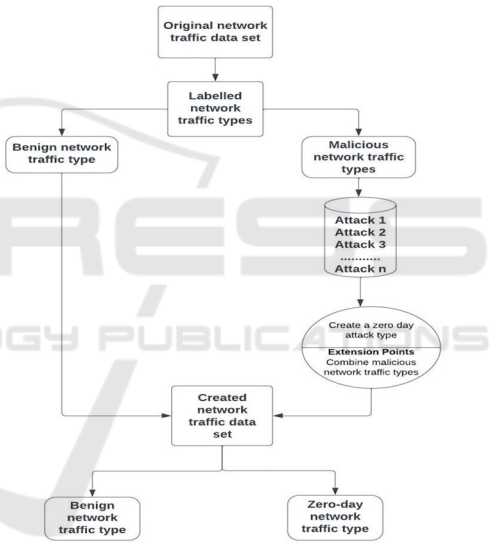

Figure 2: Approach 2.

To illustrate, Approach 2 was used to expand the IoT intrusion-2020 data set. In this data set, the benign network traffic is labelled as "normal", and malicious network traffic is labelled as "anomaly" (Ullah & Mahmoud, 2020). For the purposes of this paper, the authors created a new attack (i.e., zero-day) by combining anomaly network traffic. Specifically, a zero-day (i.e., anomaly) attack is created by combining SYN Flooding + UDP Flooding + HTTP Flooding + ACK Flooding + Host Brute force + ARP Spoofing + Scan host port + Scan port OS across the same features.

3.2 Clean the Data: Quality of Cyber Security Data for Machine Learning

The CySecML methodology include the process of ensuring that a Cyber security data set is of sufficient quality for an ML problem as part of the data preparation component (see Figure 1). The IoT intrusion-2020 (Ullah & Mahmoud, 2020) data set is examined following the requirements for data quality as reported on by Cai et al. (Cai & Zhu, 2015):

• Availability: characterise the extent to which data is easily made available and accessible to individuals who have the right level of access to obtain data. The IoT intrusion-2020 data set comply as follows to this criterium: as this is a publicly available data set, and easily accessible on the internet.

• Usability: refers to the extent in which data meets the requirements of a problem at hand. The IoT intrusion-2020 data set comply as follows to this criterium: as this data set contains benign network traffic and a variety of malicious network traffic with many instances.

• Reliability: once the question of usability has been answered, the subsequent question is whether one can trust the data in terms of accuracy and completeness amongst other factors. The IoT intrusion-2020 data set comply as follows to this criterium: given that the cyber attacks such as flooding attacks considered in this data set are real-world cyber attacks.

• Relevance: relevance of data is closely linked with usability of data with the exception that, relevance asks the question as to why this data is needed in the first place? The IoT intrusion-2020 data set comply as follows to this criterium: as this data set contains more than 80 features that include the number of packets and flow duration of benign and malicious network traffic.

• Presentation quality: refers to the manner in which data is presented in terms of description and structure amongst other factors. The IoT intrusion-2020 data set comply as follows to this criterium: this data set has clearly labelled features and network traffic types.

3.3 Feature Selection on Cyber Security Data for Machine Learning

The CySecML methodology include a step "Feature selection", see Figure 1, to identify significant features that can assist Cyber security ML models. For the case at hand, the aim is to identify significant features that are indicative of anomalous behaviour between benign and malicious zero-day network traffic types. The IoT Intrusion-2020 data set consists of benign and eight malicious network traffic type of attacks.

As part of CySecML with specific reference to the "Feature selection" step, Benford's law (BL) is presented as a feature selection method. Experiments done by the authors of this paper and previously published work (Mbona & Eloff, 2022a) highlighted the benefits of employing BL as a feature selection method to overcome the challenges relating to analysing large volumes of network traffic data sets.

4 CySecML METHODOLOGY: PROOF OF CONCEPT

This section presents implementation results of the CySecML methodology (see Figure 1). In the "Data collection or data creation" step, the IoT Intrusion-2020 data set was not only identified as a quality ML data set but it was also expanded upon to demonstrate how an existing data set can be made relevant and useful for Cyber security related ML models. The "High data quality check" step identified the data quality framework of Cai et al. (Cai & Zhu, 2015) to be a useful data quality framework for Cyber security data sets. The "Feature selection" step employed Benford's law as a simple but computationally effective method for identifying the set of significant features from the expanded IoT Intrusion-2020 data set. Based on applying these 3 steps of the CySecML methodology the following set of significant features, for the discovery of zero-day network attack types from: the IoT intrusion-2020 (Ullah & Mahmoud, 2020), UNSW-NB15 (Moustafa & Slay, 2017), CICDDOS2019 (Sharafaldin et al., 2019), and CIRACICDOHBRW2020 (Montazerishatoori et al., 2020), was identified:

• Flow duration: is defined as a measure of a set of packets passing a particular point in network traffic during a certain time interval (NetFort Technologies Limited, 2014; Umer et al., 2017). Flow features have been shown in the past to be effective for discovering malicious network traffic, see (Khodjaeva & Zincir-Heywood, 2021; Umer et al., 2017; Zavrak & Iskefiyeli, 2020).

• Flow inter-arrival times: measures the time between two consecutive network flows. The flow inter-arrival times measure was adopted by (Arshadi & Jahangir, 2017; Asadi, 2016; Sethi et al., 2020) to discover anomalous network traffic types.

• Forward and backward packets: measures packets received and sent in a network (Ullah & Mahmoud, 2020).

• Packet sizes: is a measure of actual packet size (Ullah & Mahmoud, 2020).

• Segment size: can be described as the largest amount of data in bytes that a device can handle (Ullah & Mahmoud, 2020).

• Source to destination / destination to source packet count: measures the number of packets from source to destination, and vice versa (Moustafa & Slay, 2017).

• Source/destination TCP base sequence number: can be described as a number used to keep track of every byte sent or received by a host (Moustafa & Slay, 2017).

• Number of connections of the same source/destination address: contains information about the number of connections of the same source/destination address (Montazerishatoori et al., 2020).

The above listed significant features can now be used as "Input data" in a ML model such as the IBD model for the purposes of discovering zero-day network intrusion attacks as depicted in Figure 1. The IBD model was implemented using the one-class support vector machine (OCSVM), see (Mbona & Eloff, 2022a) for model and evaluation measure details. The results in Table 1 demonstrate that "Approach 2" which is based on a combination of all malicious network traffic, produced the best results for the discovery of zero-day network traffic.

Table 1: IBD model experimental results using the IoT Intrusion-2020 data set.

Training set "Known attacks"	**Testing set "Zero-day attack"**	**F_1 score (%)**	**MCC score (%)**
Benign + ARP spoofing + Scan host port	UDP flooding	76	60
Benign + Scan port OS + Scan host port + ARP spoofing	SYN flooding + HTTP flooding	73	56
UDP flooding	Host brute force	62	31
Benign + SYN flooding + UDP flooding + HTTP flooding + ACK flooding + Host brute force	ARP spoofing + Scan host port + Scan port OS	70	59
Benign network traffic only	All combined attacks	71	64

The results in Table 1 indicate that the OCSVM algorithms performs well in terms of F_1 and MCC scores, although it performed poorly for the UDP flooding and Host brute force case. Although each case in Table 1 contains different type of attacks in the training and testing stages, i.e., "Approach 1" and "Approach 2" the results of the OCSVM are not consistent. Consequently, this research makes an important contribution to the ongoing search for designing and discovering zero-day attack scenarios.

5 CONCLUSION

This paper illustrates the convergence of Cyber security and artificial intelligence (AI) with a focus on data sets. A methodological approach (CySecML) was proposed to assist Cyber security researchers and professionals in developing Cyber security ML solutions. It is demonstrated how the CySecML methodology optimises existing techniques such as data collection or data creation, checking data quality, and identifying features. CySecML was used on several Cyber security ML projects, most notably the work performed on NIDS, as illustrated in this paper. Future work will focus on extending the data collection step by refining the methods used to generate synthetic Cyber security data and developing a metric approach for measuring the compliance to standards for data quality characteristics of Cyber security ML data. Regarding NIDS, this study makes an important contribution to the ongoing search for designing and discovering zero-day attack scenarios.

REFERENCES

Abdallah, Z. S., & Webb, G. I. (2017). Data Preparation. *Encyclopedia of Machine Learning and Data Mining*, *September 2018*, 318–327. https://doi.org/10.1007/978-1-4899-7687-1

Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. (2023). Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review*, *February*, 1–79. https://doi.org/10.1007/s10462-023-10437-z

Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, *25*, 152–160. https://doi.org/10.1016/j.jocs.2017.03.006

Amr, T. (2019). Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits. *OReilly Media*, 384.

Arshadi, L., & Jahangir, A. H. (2017). An empirical study on TCP flow interarrival time distribution for normal

and anomalous traffic. *International Journal of Communication Systems*, *30*(1). https://doi.org/10.1002/dac.2881

Asadi, A. N. (2016). An approach for detecting anomalies by assessing the inter-arrival time of UDP packets and flows using Benford's law. *Conference Proceedings of 2015 2nd International Conference on Knowledge-Based Engineering and Innovation, KBEI 2015*, *2*(6), 257–262. https://doi.org/10.1109/KBEI.2015.7436057

Bonaccorso, G. (2020). Mastering Machine Learning Algorithms: Expert techniques for implementing popular machine learning algorithms, fine-tuning your models, and understanding how they work. In *OReilly Media*. OReilly Media.

Bowles, J. K. F., Silvina, A., Bin, E., & Vinov, M. (2020). *On Defining Rules for Cancer Data Fabrication*. 168–176. https://www.ndc.scot.nhs.uk/National-Datasets/

Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery. https://books.google.co.za/books?hl=en&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=Data+preparation+for+machine+learning:+data+cleaning,+feature+selection,+and+data+transforms+in+Python&ots=Cl8GuchLpT&sig=MmuT6WgKuVEHvbXGQj91vPH2M_k

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*. https://doi.org/10.5334/dsj-2015-002

Dankar, F. K., & Ibrahim, M. (2021). Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences (Switzerland)*, *11*(5). https://doi.org/10.3390/app11052158

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *MDPI Mathematics*, *10*(15). https://doi.org/10.3390/math10152733

Financial Times. (2023, July 21). *AI systems create synthetic data to train next generation models*. https://ft.pressreader.com/v99c/20230721/281732683968639

García, S., Luengo, J., & Herrera, F. (2015). Feature selection. *Intelligent Systems Reference Library*, *72*(6), 163–193. https://doi.org/10.1007/978-3-319-10247-4_7

Kaushik, R., & Dave, M. (2021). Malware Detection System Using Ensemble Learning: Tested Using Synthetic Data. *Data Engineering and Communication Technology: Proceedings of ICDECT 2021*, *63*, 153–164. https://doi.org/10.1007/978-981-16-0081-4_16

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, *1*(October), 372–378. https://doi.org/10.1109/SAI.2014.6918213

Khodjaeva, Y., & Zincir-Heywood, N. (2021, August 17). Network Flow Entropy for Identifying Malicious Behaviours in DNS Tunnels. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3465481.3470089

Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Computer Networks*, *188*. https://doi.org/10.1016/j.comnet.2021.107840

Kim, S., Hwang, C., & Lee, T. (2020). Anomaly based unknown intrusion detection in endpoint environments. *Electronics (Switzerland)*, *9*(6), 1–21. https://doi.org/10.3390/electronics9061022

Kumar, V., & Sinha, D. (2023). Synthetic attack data generation model applying generative adversarial network for intrusion detection. *Computers and Security*, *125*. https://doi.org/10.1016/j.cose.2022.103054

Kurniabudi, Stiawan, D., Darmawijoyo, Bin Idris, M. Y. Bin, Bamhdi, A. M., & Budiarto, R. (2020). CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection. *IEEE Access*, *8*, 132911–132921. https://doi.org/10.1109/ACCESS.2020.3009843

Lu, Y., Shen, M., Wang, H., & Wei, W. (2023). Machine Learning for Synthetic Data Generation: A Review. *ArXiv Preprint ArXiv:2302.04062*. http://arxiv.org/abs/2302.04062

Mbona, I., & Eloff, J. H. P. (2022a). Detecting Zero-Day Intrusion Attacks Using Semi-Supervised Machine Learning Approaches. *IEEE Access*, *10*(July), 69822–69838. https://doi.org/10.1109/ACCESS.2022.3187116

Mbona, I., & Eloff, J. H. P. (2022b). Feature selection using Benford's law to support detection of malicious social media bots. *Information Sciences*, *582*, 369–381. https://doi.org/10.1016/j.ins.2021.09.038

Montazerishatoori, M., Davidson, L., Kaur, G., & Habibi Lashkari, A. (2020). Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech).*, 63–70. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00026

Moustafa, N., & Slay, J. (2017). The significant features of the UNSW-NB15 and the KDD99 data sets for Network Intrusion Detection Systems. *Proceedings - 2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, BADGERS 2015*, *1*(November), 25–31. https://doi.org/10.1109/BADGERS.2015.14

Mukkamala, S., & Sung, A. H. (2003). Identifying Significant Features For Network Forensic Analysis Using Artificial Intelligent Techniques. *International Journal of Digital Evidence*, *1*(4), 1–17.

NetFort Technologies Limited. (2014). Flow Analysis Versus Packet Analysis . What Should You Choose ? *White Paper*, 6.

Pu, G., Wang, L., Shen, J., & Dong, F. (2021). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology*, *26*(2), 146–153. https://doi.org/10.26599/TST.2019.9010051

Rupa Mahanti. (2019). Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. In *Quality Press*. Quality Press. https://www.proquest.com/openview/d4fd7a60ba9f04a4b438e6df4720bec3/1?cbl=34671&pq-origsite=gscholar&parentSessionId=l3HZaJC3EtvEQ49NCH9M7LjojVXg6OK878K4hWSG%2BB4%3D

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(11). https://doi.org/10.1109/TPAMI.2014.2321392

Sethi, K., Kumar, R., Prajapati, N., & Bera, P. (2020). A Lightweight Intrusion Detection System using Benford's Law and Network Flow Size Difference. *2020 International Conference on COMmunication Systems and NETworkS, COMSNETS 2020*, *10*, 1–6. https://doi.org/10.1109/COMSNETS48256.2020.9027422

Sharafaldin, I., Lashkari, A. H., Hakak, S., & Ghorbani, A. A. (2019). Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. *Proceedings - International Carnahan Conference on Security Technology*, *2019-Octob*(October). https://doi.org/10.1109/CCST.2019.8888419

Singh, U. K., Joshi, C., & Kanellopoulos, D. (2019). A framework for zero-day vulnerabilities detection and prioritization. *Journal of Information Security and Applications*, *46*, 164–172. https://doi.org/10.1016/j.jisa.2019.03.011

Ullah, I., & Mahmoud, Q. H. (2020). A Technique for Generating a Botnet Dataset for Anomalous Activity Detection in IoT Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *2020-Octob*(May), 134–140. https://doi.org/10.1109/SMC42975.2020.9283220

Umer, M. F., Sher, M., & Bi, Y. (2017). Flow-based intrusion detection: Techniques and challenges. *Computers and Security*, *70*, 238–254. https://doi.org/10.1016/j.cose.2017.05.009

Van Der Walt, E., & Eloff, J. (2018). Using Machine Learning to Detect Fake Identities: Bots vs Humans. *IEEE Access*, *6*, 6540–6549. https://doi.org/10.1109/ACCESS.2018.2796018

Zavrak, S., & Iskefiyeli, M. (2020). Anomaly-Based Intrusion Detection from Network Flow Features Using Variational Autoencoder. *IEEE Access*, *8*, 108346–108358. https://doi.org/10.1109/ACCESS.2020.3001350

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. In *O'Reilly Media*. O'Reilly Media. https://perso.limsi.fr/annlor/enseignement/ensiie/Feature_Engineering_for_Machine_Learning.pdf%0Ahttps://www.amazon.com/Feature-Engineering-Machine-Learning-Principles/dp/1491953241

Zoppi, T., Ceccarelli, A., & Bondavalli, A. (2021). Unsupervised Algorithms to Detect Zero-Day Attacks: Strategy and Application. *IEEE Access*, *9*, 90603–90615. https://doi.org/10.1109/access.2021.3090957