Dataset Selection for Attacker Group Identification Methods

Artem Pavlov, Natalia Voloshina
ITMO University
St. Petersburg, Russia
182233@niuitmo.ru, nvvoloshina@itmo.ru

Abstract— **Intrusion detection systems are an important tool for network security. Their efficiency can be improved by implementing Alert Correlation Systems. Such systems are aimed at identifying relationships between alerts themselves and between alert and properties of protected systems. One of the tasks of alert correlation systems is to identify groups of attackers. Solving this task allows to improve the accuracy of determining the threat level of malicious actors and to determine patterns of similarity between attacks. These improvements help in choosing response measures and forensic investigation. To date, there is no universal dataset suitable for testing the effectiveness of any method related to intrusion detection systems, and the most appropriate dataset for the task of attacker group identification has not been selected. The paper considers the existing approaches to the formation of requirements for datasets for use in intrusion detection tasks, and analyzes modern datasets. A list of requirements for datasets is formed for their use in testing methods for identifying groups of attackers based on the specifics of the task. Weights are determined for the requirements, and a usability rating is determined for the modern datasets. An alternative data source is proposed to meet requirements that are poorly addressed by the current datasets.**

I. Introduction

Intrusion detection systems (IDS) play an important role in the information security of modern organizations. They can improve protection against hacker attacks and malicious file campaigns. However, they tend to generate enormous number of alerts, provide heterogeneous data, create false alerts, skip real attacks and are unable to find connections between alerts [1]. To improve the performance of IDS and address some of these drawbacks, Alert Correlation Systems are used. They are aimed at identifying relationships both between alerts and between alerts and properties of protected systems. Their tasks include normalization, aggregation, correlation, false alert reduction, attack strategy analysis, alert prioritization and attack group recognition [2].

Identifying groups of attackers is aimed at discovering the similarity of the tools used in attacks and the transfer of information about the attacked system between attackers. It is also aimed at attack path reconstruction and possible impact analysis. Distinguishing groups of attackers allows to take response measures appropriate to the level of threat, to reconstruct the attack path and intentions, to identify patterns of attacks and the resources used in them.

To improve the work of IDS, algorithms based on big data and machine learning methods can be developed and applied. The applicability of the proposed methods in real-world conditions depends on the quality of the dataset on which the model is trained. Thus, algorithms based on the KDD CUP 99 dataset [3] cannot be considered verified, since significant flaws were found in the dataset [4]. Algorithms and methods based on the DEFCON CTF dataset, such as [5], have received criticism for the fact that the dataset was based on game attacks rather than real attacks [6].

Requirements for datasets for use in validating intrusion detection methods have already been proposed by various researchers. The main provisions of these requirements are discussed in the next section. However, these sets of requirements are not well suited to the task of attacker group identification. Moreover, for chronological reasons, they did not address compliance with the new datasets. These problems are discussed in this paper.

II. Dataset selection problem

A. Related work

The development of a universal dataset for testing the effectiveness of intrusion detection methods remains an unsolved problem [7]. Among the reasons for this, one can single out the nature of the data acquisition: if the data were generated by the authors, then the dataset may be unrepresentative both from the point of view of attacks and from the point of view of normal (benign) behavior. If the dataset was obtained on the basis of data from a real company, then confidentiality protection requires anonymization of much data, which leads to the loss of information useful to researchers. Moreover, collecting information about different types of attacks would require very long data recording [7].

Due to the difficulty of ensuring complete privacy of the company's data [8], some researchers prefer not to share the datasets used for testing methods in the public domain, but only describe their main features [9]. Some researchers find it difficult to perform data labeling. It is also worth noting that due to constant changes both in the structure of normal user traffic and in the types and methods of attacks, even a good dataset will lose its relevance over time [9], [10].

To overcome difficulties of developing datasets and their application, many researchers have proposed requirements that a dataset must meet for use in intrusion detection and related tasks.

Nehinbe [11] highlighted aspects such as anonymization and obtaining permission from the owner of the network, designation of the scope, data labeling, documentation of the network structure and actions of the attackers, and the absence of data gaps. Malowidzki et al. [9] identified the following features of a good dataset: it contains up to date data, is realistic, contains all types of attacks encountered in the real environment, is labeled, is correct in terms of business cycles, contains data flows. Ring et al. [12] add the requirement that the dataset should contain more normal traffic than attack data, since in real networks attacks take up a small fraction of network activity.

Ghorbani et al. [13] proposes the following set of requirements: variety of attacks, anonymization, variety of presented protocols, full coverage of the network when collecting traffic, providing complete information about the network configuration, sufficiency of the presented data characteristics, data labeling, heterogeneity, and availability of documentation. Based on this list, they presented the CICIDS-2017 and CSE-CIC-IDS-2018 datasets. Sharafaldin et al. [14] put forward an approach for selecting a dataset depending on the importance of each of these characteristics for solving a specific problem.

Ring et al. [10] propose to consider the following groups of characteristics of datasets:

- *General information:* Year of collection, public availability, presence of normal behavior, presence of attacks.
- *Nature of data*: Presence and completeness of metadata, format, method of anonymization.
- *Amount of data*: Number of streams and duration of collection.
- *Collection conditions*: Traffic type, network type, network completeness.
- *Data processing*: Division into samples, balance, data labeling.

Based on these factors, Ring et al. proposed to use the following datasets:

- CICIDS 2017
- CIDDS-001
- UGR’16
- UNSW-NB15

In this study, these datasets are used as a starting point, and newer datasets that meet the criteria above are also considered.

B. Suitable datasets

UNSW-NB15 [15]. Year of collection - 2015. The dataset contains traffic collected over 31 hours of network emulation of a small company at Cyber Range Laboratory. According to the authors of the dataset, the high quality of datasets is due to the comprehensive reflection of current threats and the inclusion of normal user behavior in the data. The IXIA Perfect Storm tool was used for generation. The first simulation generated 1 attack per second, while the second generated 10 attacks. Types of attacks include exploits, backdoors, denial of service, fuzzing and network worms. The dataset is divided into training and test samples. The total number of records is 2,540,044, 4.2% of them are attacks. 49 characteristics are presented for records.

UGR'16 [16]. The dataset is based on data received from the networks of the Spanish Internet provider for 4 months, from March to June 2016. Authors added artificially generated attack traffic. The authors also labeled the attacks they detected in the real traffic of the provider - UDP port scanning, SSH scanning, spam campaigns. IP addresses found on publicly accessible blacklists have been additionally labeled. The dataset is divided into calibration and test samples. In total, the dataset contains 16,900 million streams, 0.7% of which are attacks. There are 11 properties presented, which are based on 1-minute window network statistics.

CIDDS-001 [17]. Year of collection - 2017. To create a dataset, the OpenStack utility was used to emulate the network of a small company. The structure of the network is described in detail. The dataset contains over 32 million streams. There are 14 properties, the share of attacks in the dataset is 10.2%. The lack of a class for some records can be highlighted as the disadvantage of the dataset [18].

CICIDS 2017 [14]. Year of collection - 2017.The authors identified the main priority in the development of the kit as the generation of realistic background traffic. For this, models of behavior of 25 users were created using the protocols HTTP, HTTPS, FTP, SSH and a group of mail protocols. The network was emulated for 5 days, 6 attackers' devices and 16 victims' devices were involved. In total, the dataset contains more than 2.8 million records, 83 properties and 15 different classes of attacks. The share of attacks in the dataset is 19.7%. Among the shortcomings, the researchers highlight the presence in the dataset of more than 280 thousand records with a missing class or information, a high level of class imbalance [19], as well as the duplication of the “Fwd Header Length” property [20].

CSE-CIC-IDS2018 [13]. Year of collection - 2018. It is a continuation and expansion of the CICIDS 2017 dataset [21]. The operation of 50 attacking devices and 450 victim devices was emulated for 10 days. There are 16.2 million streams, 16.9% of which are attacks, and 83 properties of data. The data is divided into 18 classes. Some researchers called this dataset the most up-to-date for 2020, highlighted the sufficient amount of data presented and a wide range of types of attacks [22].

Dataset of intrusion detection alerts from a sharing platform [23]. Year of collection - 2019. The dataset contains intrusion data received in the SABU Alert Sharing Platform during 1 week. Data were obtained from IDS installed in 3 organizations. Nearly 12 million alerts from 34 IDS, honeypots and other sources were collected. The alerts received were significantly enriched: data on geolocation, the presence of IP in blacklists, data on DNS and network fingerprint were added. The dataset does not include the raw or processed traffic itself, so it is interesting primarily with metadata about attacks.

NF-UQ-NIDS-v2 [24]. Years of collection - 2015-2020. The authors used a combination approach, which is in line with the recommendations of Ring et al. [10]. They took UNSW-NB15, CSE-CIC-IDS2018, BoT-IoT [25] and ToN-IoT [26] datasets as a basis, extracted a piece of data from each, and processed them to a common dataset of 43 properties using *nProbe* utility. The share of attacks in the resulting dataset is 66.9%; there are 20 classes of attacks.

III. Dataset requirements for attacker group recognition task

The task of identifying groups of attackers differs from the task of identifying intrusions, for which the above requirements for the datasets were formed. Intrusion detection methods are aimed at defining if the stream is malicious or not, or to define an attack class the stream belongs to. The attacker group identification in terms of intrusion detection systems can be viewed as clustering of alerts of defensive systems according to the following [2]:

- *The similarity of the network fingerprints of the tools used and attacks performed (Fig. 1):* Different malicious actors may use custom attack tools and payloads even for the same vulnerabilities. Different attackers also tend to have various working hours, attack source location and action patterns.

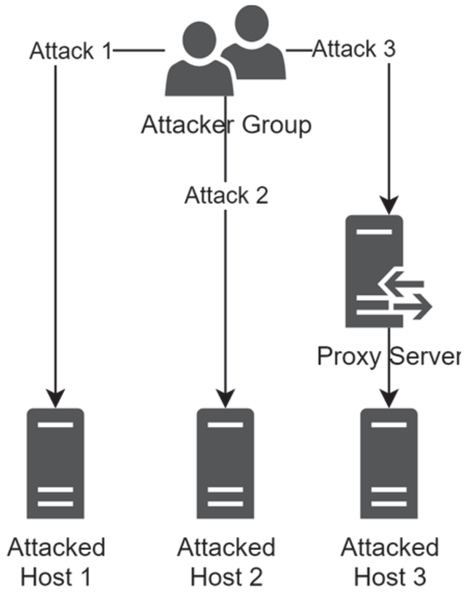

Fig 1. Different host attack scenario

- *Correspondence of alerts to different stages of known single attack path (Fig. 2)*: Some alerts may refer to the same attack, which takes more than one request or stream to perform data exfiltration or command execution.

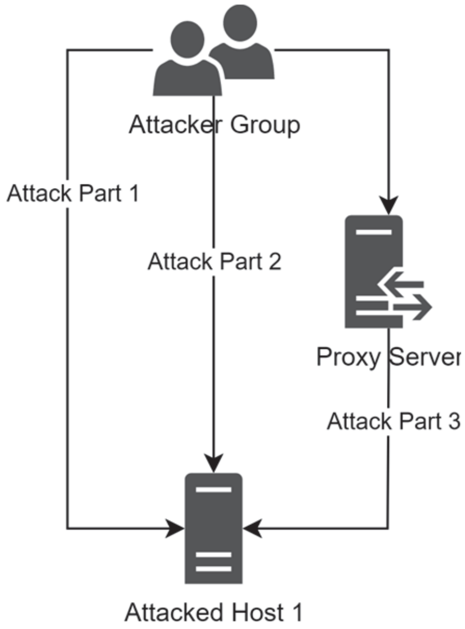

Fig. 2. Attack path scenario

- *Compliance of alerts and groups of alerts with the asset impact scheme or prerequisites/consequences model (Fig. 3)*: Various data, credentials or access gathered during the attack process have limited applicability for further attack steps. Discovering which assets attacker could exploit with the known access level is critical for defensive measure selection.

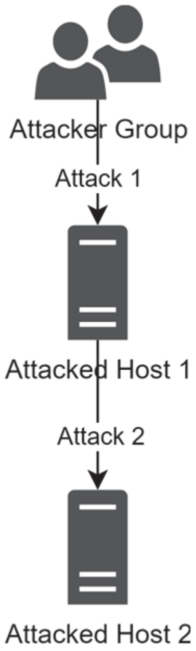

Fig. 3. Asset impact attack scenario

As we identify the relationships between the discovered actions of the attackers, the presence, quantity and quality of normal traffic in the dataset does not affect its applicability in the task of identifying groups of attackers. The task only requires attack data to be solved.

The requirement to balance the classes of attacks is becoming urgent. The dataset is considered unbalanced when the ratio of the number of elements of the majority to the minority class is more than 100:1 [27]. Despite the possibility of training on such datasets using the technique of splitting the majority class, duplicating elements of minority or other approaches, the quality of the resulting model decreases [28].

The presence of complex, multi-stage attacks in the dataset becomes necessary, since within the framework of defining the groups of attackers, security alerts are linked along the stages of the attack path. Real-world APT attack usually requires several steps to be successful [29].

The changes made to the original data should not affect the ability to separate the sources of attacks, restore the sequence of stages of attacks, so this becomes an additional requirement to the anonymization process.

The final set of requirements for the use of datasets in the task of identifying groups of attackers, based on the suggestions of Ghorbani et al. [13] and the comments above, is suggested as the following:

1) *Variety and relevance of attacks:* Attacks seen in real networks, common vulnerabilities and attack classes change over time. The types of attacks presented in the dataset must correspond to those that were actual at the time of collection. They should relate to different stages of attack campaign, represent the use of different tools, different types of vulnerabilities. Detection evasion and payload obfuscation techniques should also be present.

2) *Attack type balance:* The number of presented attacks by type should be sufficient to highlight the features of this type, but not high enough to cause overfitting. To achieve this, class splitting technique can be applied.

3) *Large attack number:* The dataset should contain a significant number of attacks, sufficient to highlight the characteristic features of different attack vectors.

4) *Presence of multi-stage (complex) attacks:* Attacks that exploit the results achieved by previous attacks by the attackers are required to be present in the dataset, as this behavior is common in real attacks. This should include privilege escalation and lateral movement.

5) *Correct anonymization:* The output should not allow obtaining confidential information about the network in which the collection was carried out and about the nodes interacting with it. However, anonymization should not affect the ability to separate attack sources and reestablish the sequencing of complex attacks. It should remain possible to collect metadata about the attacking hosts, like geolocation, ASN number, DNS fingerprint should remain, either these data may be provided in the dataset in advance.

6) *Protocol variety*: If the dataset is of artificial origin, then the protocols presented in it according to statistical characteristics, for example, frequency, must correspond to the data of a real network that the researchers emulate.

7) *Full network coverage:* The dataset should represent all interactions within the network, including broadcast traffic, DNS requests, UDP and ICMP communication.

8) *Complete documentation:* The description of the dataset should provide information about the endpoints, their operating systems, installed software, network connectivity. The description of the collection techniques and conditions, network configuration, collection periods and methods used in processing must be presented.

9) *Feature sufficiency:* The features present in the dataset should be as close as possible to the data that could be extracted directly from the traffic itself. With different approaches to analysis, different properties can be shown to be effective.

10) *Correct data labeling:* The classes of attacks must be correctly set for each stream present in the dataset. There should be no gaps in data features among records. No duplicates in records or features should be present.

IV. Dataset selection

Now we consider the compliance of datasets with the above requirements. For a general assessment, we use a framework based on the proposal by Sharafaldin et al. [14]. The dataset score is calculated using the following formula:

$$Score = \sum_{i=1}^{n} W_i F_i$$

Where n is the number of considered requirements for the dataset, W_i is the weight of the requirement in solving a specific problem, and F_i is the value of the characteristic of meeting the requirement in a specific dataset.

For the 10 requirements under consideration in the problem of determining the groups of attackers, the weights are determined by the following vector:

W = [0.2; 0.1; 0.1; 0.1; 0.15; 0.025; 0.025; 0.05; 0.15; 0.1]

The greatest weight is given to the variety and relevance of attacks, since they determine the ability to link alerts in real conditions. There should be no type of attack, the features of which allow bypassing link detection mechanisms.

Correct anonymization and feature sufficiency determine our ability to obtain the maximum useful data for analysis, therefore their weight is also high.

Attack type balance, large attack number and correct labelling affect the quality of the resulting models in the case of using machine learning algorithms, which is important, but not critical. Complex attack presence affects the ability to disclose one of the attacker group scenarios, which also makes its influence notable, yet not crucial.

Other factors have a certain influence on the level of applicability in the problem, but, in the case of high dataset quality in terms of other requirements, their influence is not too high.

The values for characteristics were obtained either from the dataset technical papers or from the datasets themselves using the *pandas* [30] utility for analysis.

All considered datasets contain an actual set of attacks, therefore, a variety of classes is considered for evaluation. Blacklist, Spam classes are not considered as attack classes. The “Correct labeling” score for CIDDS-001 and CICIDS 2017 has been downgraded due to known data issues.

For characteristics defined by a number, the result of each dataset is defined as the ratio of the characteristic value in the dataset to the maximum value among all datasets:

$$F_i = \frac{Val_i}{Val_{MAX}}$$

For Requirement 1Val_{MAX} is 20 (NF-UQ-NIDS-v2 dataset), for Requirement 3 – 50822681 (NF-UQ-NIDS-v2 dataset), for Requirement 9 – 83 (CICIDS 2017 and CSE-CIC-IDS2018).

To determine the balance of the dataset (Req. 2), the Shannon Equitability Index is used:

$$Balance = \frac{-\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n}}{\log k}$$

TABLE I. DATASET SCORES

Requirement	Req. Weight	Dataset Score						
		UNSW-NB15	**UGR'16**	**CIDDS-001**	**CICIDS 2017**	**CSE-CIC-IDS2018**	**Sharing Platform**	**NF-UQ-NIDS-v2**
Variety and relevance of attacks	0.2	0.40	0.25	0.20	0.75	0.90	0.40	1.00
Attack type balance	0.1	0.724	0.694	0.235	0.568	0.825	0.278	0.477
Large attack number	0.1	0.00208179	0.35417384	0.06408938	0.01097238	0.05597243	0.23231407	1
Presence of complex attacks	0.1	0	0	0	1	1	0	1
Non-destructive anonymization	0.15	1	1	1	1	1	0	1
Protocol variety	0.025	1	1	1	1	1	1	1
Full network coverage	0.025	1	0	1	1	1	1	1
Complete documentation	0.05	1	0	1	1	1	0	1
Feature sufficiency	0.15	0.59	0.13	0.17	1.00	1.00	0.48	0.52
Correct data labeling	0.1	1	1	0.9	0.9	1	1	1
Total Score		0.5912	0.45	0.4352	0.7979	0.8681	0.3533	0.8754

Where n is the total number of records in the dataset, k is the number of classes, c_i is the number of records in class i. Since the problem of alert clustering is being solved, the total number of records is considered as the number of attacks.

In Table I, the data on the values of the characteristics and the final score for each of the considered datasets are presented. Datasets NF-UQ-NIDS-v2, CSE-CIC-IDS2018 and CICIDS 2017 have the best overall scores.

V. DISCUSSION AND CONCLUSIONS

The paper considers various approaches to formulating requirements for intrusion detection datasets. A brief overview of modern datasets in this area is given. Attacker group recognition scenarios are presented. For the task of detecting groups of attackers, the following requirements for datasets are formed:

- Variety and relevance of attacks
- Attack type balance
- Large attack number
- Presence of complex attacks
- Non-destructive anonymization
- Protocol variety
- Full network coverage
- Complete documentation
- Feature sufficiency
- Correct data labeling.

Weights are specified for the requirements. Based on the dataset selection framework, the following datasets are recommended for attacker group detection methods efficiency verification:

1) NF-UQ-NIDS-v2
2) CSE-CIC-IDS2018
3) CICIDS 2017

Other mentioned datasets have notably lower score for the use in attacker group recognition methods.

Despite the presence of complex attacks in some datasets, the share of such records is small. They will not reflect the full ability of the attackers to exploit the results achieved earlier. Moreover, the datasets do not include techniques for bypassing protective measures and payload obfuscation.

Cyber polygons such as StandOff [31] could become an alternative source of data that solves the indicated problems. During the competition various attacking teams perform actions aiming to compromise systems and trigger business risks in a network of companies, for the security of which the team of defenders is responsible. Attackers need to enter the internal network of the organization through the Demilitarized Zone or by means of social engineering, bypass the means of detection and monitoring. Defenders are allowed to take actions to prevent attackers from gaining further control and even eliminate them from the hosts.

Therefore, the data from such platforms would be similar to those that could be obtained from real-life malicious campaigns. Moreover, the absence of confidential data in the networks of the emulated companies would help to avoid the need for anonymization and, at the same time, the potential loss of useful data for attacker group identification.

Other approaches to solving the problem of complex attacks and evasion techniques presence may be based on known

Advanced Persistent Threat strategy emulation and Red Team traffic recording.

For creating industry-specific attack datasets, digital twins may be used. They are based on creating the digital representation of vital parts of the network, which allows performing attacks and improving security mechanisms [32] without causing threat to real systems.

Future work may focus on implementing methods for identifying attacker groups using the datasets described above. These methods should implement one or more of scenarios mentioned. Modern asset discovery systems may aid in development of methods, which rely on the data about systems they protect.

Researches in the area of Intrusion Detection dataset development may also focus on creating datasets based on continuous real-world traffic monitoring. This requires creating automated privacy protection and anonymization methods. Country and multiple-country level is preferred for data capture to be able to gather enough intel. Government SOC (Security Operation Center) or CERT (Computer Emergency Response Team) are suitable for this. Even though the implementation of this approach might look like a hard task, the successful realization can overcome the loss of relevance over time, which is an issue for all modern IDS datasets.

ACKNOWLEDGMENT

This work was supported by grant 620164 “Artificial Intelligence Methods for Cyber-Physical Systems”. Authors wish to thank the anonymous reviewers for their valuable comments and feedback that helped improve the quality of the paper.

REFERENCES

[1] S.A. Mirheidari, S. Arshad, R. Jalili, “Alert Correlation Algorithms: A Survey and Taxonomy”, Cyberspace Safety and Security, Lecture Notes in Computer Science, 2013, vol. 8300, pp. 183-197.

[2] A. Pavlov, N. Voloshina, “Analysis of IDS Alert Correlation Techniques for Attacker Group Recognition in Distributed Systems”, NEW2AN, Lecture Notes in Computer Science, 2020, vol. 12525, pp. 32-42.

[3] KDD Cup 1999, Web: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[4] M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani “A Detailed Analysis of the KDD CUP 99 Data Set”, 2009 IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2009, pp. 1-6.

[5] O. Dain, R.K. Cunningham, “Fusing a heterogeneous alert stream into scenarios”, Proceedings of the 2001 ACM Workshop on Data Mining for Security Applications, Philadelphia, PA, 2001, pp. 1–13.

[6] R. Smith, N. Japkowicz, M. Dondo, P. Mason, “Using unsupervised learning for Network Alert Correlation”, Advances in Artificial Intelligence, Canadian AI 2008, Lecture Notes in Computer Science, 2008, vol. 5032, pp. 308-319.

[7] R. Koch, M. Golling, G.D. Rodosek, “Towards comparability of intrusion detection systems: New data sets”, Proceedings of the TERENA Networking Conference, 2014, vol. 7.

[8] R. Pang, M. Allman, V. Paxson, J. Lee, “The devil and packet trace anonymization”, SIGCOMM Comput. Commun. Rev., 2006, vol. 36, pp. 29–38.

[9] M. Malowidzki, P. Berezinski, M. Mazur, “Network Intrusion Detection: Half a Kingdom for a Good Dataset”, NATO STO SAS-139 Workshop, Portugal, 2015.

[10] M. Ring, S. Wunderlich, D.A. Scheuring, “A Survey of Network-based Intrusion Detection Data Sets”, Computers & Security, 2019, vol. 86, pp. 147-167.

[11] J.O. Nehinbe, “A critical evaluation of datasets for investigating IDSs and IPSs Researches”, IEEE 10th International Conference on Cybernetic Intelligent Systems (CIS), 2011, pp. 92-97.

[12] M. Ring, S. Wunderlich, D. Gruedl, D. Landes, “Creation of Flow-Based Data Sets for Intrusion Detection”, Journal of Information Warfare, 2017, vol. 16, issue 4, pp. 40-53

[13] I. Sharafaldin, A. Habibi Lashkari, A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization”, Proceedings of the 4th International Conference on Information Systems Security and Privacy – ICISSP, 2018, pp. 108-116.

[14] I. Sharafaldin, A. Gharib, A. Habibi Lashkari, A. Ghorbani, “Towards a Reliable Intrusion Detection Benchmark Dataset”, Software Networking, 2017, vol. 1, pp. 177-200.

[15] N. Moustafa, S. Jill, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)”, Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6.

[16] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, “UGR‘16: A new dataset for the evaluation of cyclostationarity-based network IDSs”, Computers & Security, 2018, vol. 73, pp. 411-424.

[17] M. Ring, S. Wunderlich, D. Gruedl, D. Landes, “Flow-based benchmark data sets for intrusion detection”, Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS), 2017, pp. 361-369.

[18] A. Verma, V. Ranga, “On Evaluation of Network Intrusion Detection Systems: Statistical Analysis of CIDDS-001 Dataset Using Machine Learning Techniques”, Pertanika Journal of Science and Technology, vol. 26, pp. 1307-1332.

[19] R. Panigrahi, S. Borah, “A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems”, International Journal of Engineering & Technology, 2018, vol. 7, pp. 479-482.

[20] K. Kurniabudi, S. Deris, D. Darmawijoyo, “CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection”, IEEE Access, 2017, pp. 1-12.

[21] A. Thakkar, R. Lohiya, “A Review of the Advancement in Intrusion Detection Datasets”, Procedia Computer Science, 2020, vol. 167, pp. 636-645.

[22] J. Leevy, T. Khoshgoftaar, “A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data”, Journal of Big Data, 2020, vol. 7, pp. 1-19.

[23] M. Husák, M. Žádník, V. Bartoš, “Dataset of intrusion detection alerts from a sharing platform”, Data in Brief, 2020, vol. 33, pp. 1-12.

[24] M. Sarhan, S. Layeghy, N. Moustafa, “Towards a Standard Feature Set of NIDS Datasets”, eprint arXiv:2101.11315, 2021, pp. 1-13.

[25] N. Koroniotis, N. Moustafa, E. Sitnikova, “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: BoT-IoT dataset”, Future Generation Computer Systems, 2019, vol. 100, pp. 779-796.

[26] N. Moustafa, A. Alsaedi, Z. Tari, “TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems”, IEEE Access, 2020, vol. 8, pp. 165130-165150.

[27] H. He, E. Garcia, “Learning from imbalanced data”, IEEE Transactions on Knowledge and Data Engineering, 2009, vol. 21, num. 9, pp. 1263-1284.

[28] P. Kang, S. Cho, “EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems”, Proceedings of the International Conference on Neural Information Processing. Hong Kong, China, 2006, pp. 837–846.

[29] J. Navarro, P. Parrend, A. Deruyver, “A Systematic Survey on Multi-step Attack Detection”, Computers & Security, 2018, vol. 76, pp.214-249.

[30] Pandas, Web: https://pandas.pydata.org/.

[31] The StandOff, Web: https://standoff365.com/.

[32] Myllyla, J., Costin, A., “Reducing the Time to Detect Cyber Attacks: Combining Attack Simulation With Detection Logic”, S. Balandin, Y. Koucheryavy, & T. Tyutina (Eds.), FRUCT '29: Proceedings of the 29th Conference of Open Innovations Association FRUCT, 2021, pp. 465-474.