Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 636–645

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

A Review of the Advancement in Intrusion Detection Datasets

Ankit Thakkar, Ritika Lohiya*

[a]*Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad - 382 481, Gujarat, India.*

Abstract

The research in the field of Cyber Security has raised the need to address the issue of cybercrimes that have caused the requisition of the intellectual properties such as break down of computer systems, impairment of important data, compromising the confidentiality, authenticity, and integrity of the user. Considering these scenarios, it is essential to secure the computer systems and the user using an Intrusion Detection System (IDS). The performance of IDS studied by developing an IDS dataset, consisting of network traffic features to learn the attack patterns. Intrusion detection is a classification problem, wherein various Machine Learning (ML) and Data Mining (DM) techniques applied to classify the network data into normal and attack traffic. Moreover, the types of network attacks changed over the years, and therefore, there is a need to update the datasets used for evaluating IDS. This paper list the different IDS datasets used for the evaluation of IDS model. The paper presents an overview of the ML and DM techniques used for IDS along with the discussion on CIC-IDS-2017 and CSE-CIC-IDS-2018. These are recent datasets consisting of network attack features and include new attacks categories. This paper discusses the recent advancement in the IDS datasets that can be used by various research communities as the manifesto for using the new IDS datasets for developing efficient and effective ML and DM based IDS.

Keywords: Intrusion Detection System; Intrusion Detection Datasets; Attack Classification; Performance Evaluation; Machine Learning Techniques; CIC-IDS-2017 dataset; CSE-CIC-IDS-2018 dataset

1. Introduction

According to the statistics reported for Cyber Security, the damages caused by the cyber attacks are expected to reach up to three trillion by 2021 with the probability of executing zero-day exploits one-per-day [1]. Moreover, the amount of the information stored in private as well as public clouds operated by data-driven companies such as Amazon Web Services, Facebook, and Twitter, will be increased hundred times by 2022 [2]. Thus, an increase in the data demand for more proficient security systems. The computer systems with loopholes, security mechanisms with incompetent security policies, and lack of knowledge about the attacks and crimes have increased the targets for placing attacks in the network. Network attacks such as ransomware, identity theft, data theft, denial of service, and

* Corresponding author.
E-mail addresses: ankit.thakkar@nirmauni.ac.in (Ankit Thakkar)., 18ftphde30@nirmauni.ac.in (Ritika Lohiya).

10.1016/j.procs.2020.03.330

Table 1. Comparison of Different Intrusion Detection Systems [4]

Signature-based	Anomaly-based
Effective in identifying known attacks by performing contextual analysis	Detects unknown attacks and vulnerabilities along with known attacks.
It depends on the system software and operating system for identifying the attacks and vulnerabilities.	It is less dependent on the operating system, rather examines the network patterns for identifying attacks.
The attack signature database should be updated regularly.	It builds profiles of the observed network communication for identifying the attack patterns.
The signature-based IDS has a minimum knowledge of protocols.	The anomaly-based IDS performs protocol analysis to study the packet details.

zero-day attacks are difficult to trace using the standard security mechanisms such as firewall and anti-virus software [3]. Therefore, an Intrusion Detection System (IDS) used to examine the information flowing through the network and to generate an alarm for the probable malicious activities generated by the intruders. An IDS detects the intrusions either by extracting the signatures from the network packets or by analyzing the attacks patterns. An IDS that detects the intrusion by studying the signatures is termed as Signature-based IDS. The Signature-based IDS generates an alert for the matched signature patterns stored in the signature database. In contrast, an IDS detecting attacks based on the attack patterns are referred to as Anomaly-based IDS. A comparison table of the different IDS is presented in Table 1.

Regardless of the type of IDS, the basic architecture of IDS consists of four steps as shown in Figure 1. The network packets are captured using network sensors or network sniffing tools. The captured data is then filtered and examined. The filtering is performed based on filtering rules, and then signature patterns are matched with the already available signature database. An alert is generated by the IDS when a match is found with the stored signature database.

The evaluation of an IDS model can be performed by implementing Machine Learning (ML) and Data Mining (DM) techniques to classify the network traffic into benign and malicious traffic flow. The ML and DM techniques implemented on the IDS datasets contains labeled data and network traffic features. These help the classifier to learn different attack patterns to detect a particular attack. The features of the dataset help the classifier to learn the normal traffic patterns as well as attack patterns through which the classifier is able to classify the input data [5]. The dataset used for training the classifier is built by monitoring the network traffic for a particular interval of time. The dataset consists of normal network traffic and anomalous network traffic that helps the classifier to identify the patterns of the data with a sufficient amount of examples. The data collected is divided into a training set and test set for training and testing the classifier, respectively. Thus, various ML and DM techniques used for developing an IDS [6].

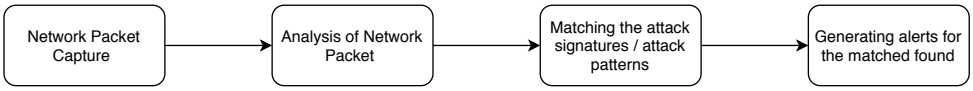

Fig. 1. Architecture of IDS

The IDS datasets consist of labels derived from observing the patterns of the network traffic data, and therefore, these datasets do not work well with zero-day exploits [7]. Along with the dataset, applying an appropriate technique for classification of attacks is also important. There are multitude of varied techniques that have been successfully used for IDS [8], [9], [10]. However, each of the algorithm used for IDS trains and tests the dataset in a different manner. The algorithms used for IDS implemented on DARPA [11], KDD CUP 99 [12], or NSL-KDD dataset [13] having network instances grouped as train set and test set. The efficiency of the developed system can be tested and compared based on factors such as parameter optimization, feature optimization, and variability in the size of the dataset.

Apart from the performance of the available datasets and the techniques used for IDS, choosing an appropriate performance metric is also one of the crucial factors that should be taken into consideration. The most commonly used metrics for showing the effectiveness of the system is accuracy [14]. The accuracy is calculated by considering the small portion of the test set or the average accuracy across many test sets is evaluated, accuracy for a specific category of attacks is measured, or the accuracy of correctly classified samples from the training set is presented

[15], [16], [17]. Therefore, a single performance metric is not sufficient to measure the efficiency of the algorithm. It is necessary to consider confusion matrix and find the number of false positives and false negatives to derive other performance metrics such as Detection Rate (DR), False Positive Rate (FPR), precision, and recall [13]. The accuracy of a particular attack type is also a critical aspect as the classifier may give better accuracy for one attack type but may fail for classifying the other [18].

The contribution of the paper can be summarized as:

- Overview of the ML and DM techniques used for IDS.
- Review of the intrusion detection datasets for performance evaluation of IDS.
- Discussion on CIC-IDS-2017 and CSE-CIC-IDS-2018 with characteristics and limitations of the datasets.

The roadmap of the paper is as follows: section 2 discusses the techniques for IDS. Section 3 is a discussion on IDS datasets and section 4 presents a study on the datasets CIC-IDS-2017 and CSE-CIC-IDS-2018. We have concluded the paper with future research scope in section 5.

2. Techniques for Intrusion Detection System

Various ML and DM methods implemented for developing an IDS are shown in Figure 2. For instance, an overview of ML methods used for IDS is presented in [19]. The paper describes different hybrid and ensemble methods along with feature selection techniques. The literature survey highlights the homogenous and heterogeneous ensemble methods implemented for IDS. A survey of ML and DM methods presented in [6], wherein the paper discusses the algorithms implemented for IDS using different subsets of KDD CUP 99 [12] dataset.

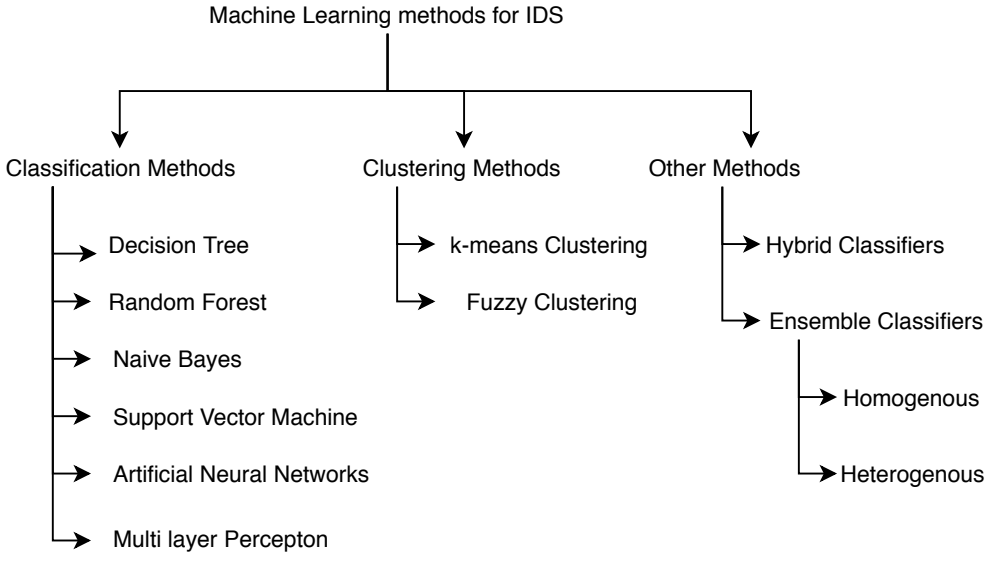

Fig. 2. Taxonomy of techniques used for IDS

A study on NSL-KDD dataset discussed in [13], where ML algorithms implemented on NSL-KDD dataset [12]. However, the study focuses on measuring the efficiency of the dataset by implementing ML algorithms. The experiments were conducted using the WEKA tool [20] and performance of the dataset for different classifier was recorded. The paper concluded that it is not necessary to consider all the features for training and classification of attacks and NSL-KDD is clearly a refined version of KDD CUP 99 dataset [13]. A similar kind of issue was addressed in [5] where common ML algorithms were surveyed, and challenges to compare the efficiency of various techniques were stated for KDD CUP 99 dataset. The paper stated that a major concern to perform the comparative analysis is the lack of an effective testbed. The paper listed the common issues with KDD CUP 99 dataset and suggested the use of

the same size of samples of the same dataset for performing a comparative analysis of different attack classification techniques.

A set of guidelines has been introduced in [7] to bridge the gap between the current requirements of IDS datasets and their shortcomings. The paper also discusses techniques to build datasets using these guidelines. Apart from publicly available datasets, simulated datasets have also been used for measuring the performance of IDS in wired and wireless networks. In [21], a simulated dataset was created for mobile ad-hoc networks and ML algorithms were applied to detect anomalies in the wireless networks.

A hybrid approach is proposed in [16] by combining Naïve Bayes (NB) classifier with Feature-Vitality based reduction method. The simulations were performed on NSL-KDD dataset and a reduced set of features was used for attack classification. The NSL-KDD dataset consists of 41 features that were reduced to 25 features by using the feature reduction method. The proposed method showed 98% of accuracy. Unsupervised ML techniques such as *k*-means clustering is used in [22] to analyze the NSL-KDD dataset. Here, 20% of the instances of NSL-KDD dataset grouped into four clusters.

Apart from supervised and unsupervised learning methods, semi-supervised learning also used for intrusion detection. A fuzzy-based semi-supervised ML method proposed in [23], wherein unlabeled instances are considered for performance evaluation. Here, a single-layer feed-forward neural network is used to train the model to deliver a fuzzy membership vector to learn each attack category separately. The experiments performed on NSL-KDD dataset reveals that unlabeled instances with low and high fuzziness value contribute more to improve the accuracy of the system.

Ensemble classifiers are also being used for attack classification. An ensemble approach for binary as well as multi-class classification is presented in [24]. It is based on a greedy randomized adaptive search combined with random forest classifier. The proposed approach used a greedy randomized adaptive search procedure for building randomized decision trees. The experiments performed on NSL-KDD dataset by using three feature selection methods namely, Information Gain (IG), Symmetrical Uncertainty (SU), and Correlation-based Feature Selection (CFS) method [24]. The results depicted that the proposed approach outperformed ML techniques like random forest classifier, multilayer perceptron, and NB and feature selection methods improved the accuracy of the system.

Extreme ML classifiers are also used in studying the network traffic profiles for classifying attacks. Network traffic profiles are examined in [25] where online sequential extreme ML applied for detecting intrusions. The proposed framework utilizes alpha profiling for minimizing the computational time, beta profile for minimizing the training dataset, and features are reduced by using the combined approach of filter, correlation, and consistency. The experiments conducted on NSL-KDD dataset that gave 98.6% accuracy with 1.7% FPR for binary classification and 97.6% accuracy and 1.7% FPR for multi-class classification. The experiments also performed on Kyoto University dataset yielding 96.37% accuracy and 5.7% FPR. A summary of the techniques used for IDS is presented in Table 2.

3. Review of Intrusion Detection Datasets

An intrusion detection dataset can be developed by collecting information from varied sources such as network traffic flows that contains information about the host, user behavior, and system configurations [26]. This information is required to study the attack patterns and abnormal activity of various network attacks. The network activity is collected through a router or network switch. After collecting the incoming and the outgoing network traffic, network flow analysis is performed to study the network traffic. Flow analysis can be described as the process of analyzing the network packet information such as source IP address, destination IP address, source port number, destination port number, type of network services to name a few [27]. The network host delivers the system configurations and user information that cannot be extracted from the network flow analysis. For instance, information obtained through failed login attempts by observing the intrusion activity.

Network security analysts can detect intrusion by observing the information obtained from network packets through network flow analysis. A few attack categories as described in DARPA dataset [11] can be listed as follows:

- Denial of Service (DoS): It is an intrusion attack performed by making the network resources busy and unavailable to the legitimate users.
- User to Root (U2R): It is an intrusion attack caused by hampering the authenticity of the user caused by permitting the root access to the intruder.

Table 2. Summary of the Techniques for IDS

Ref	Technique	Feature Selection	Dataset	Results Analysis
[16]	NB	Feature Vitality-based	NSL-KDD	A hybrid approach is proposed by combining NB with feature vitality based feature selection method to achieve accuracy of 98%
[22]	k-means	Not used	NSL-KDD	The instances of the dataset are grouped into clusters representing four attack classes along with normal traffic.
[13]	J48, SVM, NB	Correlation-based	NSL-KDD	The performance of the classifier is improved by using feature selection method. Out of the three classifiers used J48 outperforms in terms of classification accuracy (99.8%).
[25]	Sequential extreme ML	Combined approach on filter, correlation, and consistency	NSL-KDD, Kyoto	The experiments are conducted on NSL-KDD gives 98.6% accuracy for binary classification and 97.6% accuracy for multi-class classification. The experiments performed on Kyoto dataset yields 96.37% accuracy.
[24]	Greedy Adaptive Randomized Forest	IG, SU, and CFS	NSL-KDD	The experimental results are presented for binary as well as multi-class classification. The results show that the proposed approach achieves the highest accuracy with SU based feature selection method for both binary (85.05%) and multi-class (77.6%) classification compared to other methods.
[23]	Fuzzy Neural Network	Not used	NSL-KDD	The training set is divided into low and high fuzziness groups and fuzzy membership vector is used to study the instances. The proposed model gives the accuracy rate of 84.12%

- Remote to Local (R2L): It is an intrusion attack caused by breaking the integrity of the network and permitting the local network access to the intruder.
- Probe: It is an intrusion activity performed by scanning the network and gathering all network-related information about the network activities carried out in the network.

The intrusion detection datasets generated from the real network traffic traces are presented in Table **??**. These datasets were used for the performance evaluation of IDS by many researchers. To name a few, the first intrusion detection dataset was created by the MIT Lincoln Laboratory in 1998 and it was named as DARPA under the DARPA funded project [11]. Later in 1999, the tcpdump files of DARPA was refined and processed by the researchers of the University of California to form KDD CUP 99 dataset [12]. The KDD CUP 99 dataset was formed with a large number of duplicate and redundant records which were removed to form NSL-KDD dataset [13]. To evaluate the IDS alert correlation techniques a dataset was created by capturing the flag details of the network packets. This dataset was named as DEFCON and defined attack categories such as port scanning and buffer overflow [28]. CAIDA and LBNL IDS datasets are formed by examining the traces of the network flow and network packets. CAIDA was developed by the Center of Applied Internet Data Analysis [29] and LBNL was developed by Lawrence Berkeley National Laboratory [30]. The united states military academy generated a dataset named CDX based on the network warfare competition and this dataset was used to evaluate the IDS alert rules [31]. The Kyoto dataset [32] and Twente dataset [33] were developed by analyzing the activities of honeypots deployed in their respective university areas. UMASS dataset was formed by examining the trace files of the network packets and wireless applications [34] and ISCX IDS 2012 by observing the alpha and beta profile of the network packets [7] whereas AFDA dataset consists of features showing attack pattern and system call traces [35]. CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets are

the recent datasets that were developed by the Canadian Institute of Cyber Security, by performing intrusion traffic characterization. These datasets consist of seven attack categories that describe the recent attack scenarios [36].

Most of the research performed in the field of IDS has used KDD CUP 99 and NSL-KDD for measuring the performance of different classifiers [13]. These datasets consist of 41 features and four attack categories. The extracted features are categorized into four classes as follows:

- Basic: These features extracted from the open TCP/IP connection established between the communicating parties.
- Host: These features extracted from the host connections based on the protocol used and type of service.
- Traffic: These features extracted from the connections having the same service.
- Content: These features extracted from the sequential patterns of the data to detect intrusions.

The KDD CUP 99 and NSL-KDD dataset lists a few attack categories while CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets list a new range of attacks generated from real network traffic features such as Distributed Denial of Service, Denial of Service, brute force, XSS, SQL Injection, Botnet, Web attack, and Infiltration. These datasets have labeled instances with more than 80 features.

Table 3: Summary of IDS Datasets generated from the real network traffic traces

Dataset Name	Developed By	Features	Attack types	Description
DARPA	MIT Lincoln Laboratory	41	Dos, R2L, U2R, Probe	It does not represent real network traffic, absence of false-positive instances, irregularities in attack data instances.
KDD CUP 99	University of California	41	Dos, R2L, U2R, Probe	It consists of redundant and duplicate data samples.
NSL-KDD	University of California	41	Dos, R2L, U2R, Probe	Refined version of KDD CUP 99 dataset and consist of a limited number of attack types.
DEFCON	Shmoo Group	Flag traces	Telnet Protocol Attacks	Features are captured through the ”Capture the Flag” competition.
CAIDA	Center of Applied Internet Data Analysis	20	DDoS	It consists of instances that are very specific to a particular kind of attack or internet activity.
LBNL	Lawrence Berkeley National Laboratory	Internet traces	Malicious traces	It consist of 100 hours of activity specifying the traces of packet header for identifying malicious traffic.
CDX	United States Military Academy	5	Buffer Overflow	This dataset utilized network tools Nikto and Nessus to capture the traffic and was used to evaluate the IDS alert rules.
Kyoto	Kyoto University	24	Normal and Attack sessions	It was developed by deploying honeypots in the network but do not describe any details about the attack types.
Twente	Twente University	IP flows	Malicious traffic, Side-effect traffic, Unknown traffic, and Uncorrelated alerts	The size of the dataset is small and scope of attack types is limited.
ISCX2012	University of New Brunswick	IP flows	DoS, DDoS, Brute-force, Infiltration	This dataset consist of network scenarios with intrusive activities and labeled data instances.

Continued on next page

Table 3 – *Continued from previous page*

Dataset Name	Developed By	Features	Attack types	Description
AFDA	University of New South Wales	System call traces	Zero-day attacks, Stealth attack, C100 Webshell attack	This dataset consists of 10 attacks vectors along with the traces of the other data instances but has a limited range of attacks.
CIC-IDS-2017	Canadian Institue of Cyber Security	80	Brute force, Portscan, Botnet, Dos, DDoS, Web, Infiltration	Network profiles are used to generate the dataset in a specific manner.
CSE-CIC-IDS-2018	Canadian Institue of Cyber Security	80	Brute force, Portscan, Botnet, Dos, DDoS, Web, Infiltration	Network profiles are used to generate the dataset in a specific manner.

4. Discussion on Recent IDS Dataset: CIC-IDS-2017 and CSE-CIC-IDS-2018

The behavioral patterns of network attacks change gradually and therefore, it is required to upgrade the conventional datasets in the dynamic environment. This will help in manifesting different network traffic scenarios and attack patterns that are easy to adapt, learn and redefine [37]. Moreover, choosing an appropriate dataset is also a critical task. Some datasets are developed by specific organizations for their research purpose and not publicly available whereas the datasets that are publicly available contains records that may not match the current technological demand. In fact, these publicly available datasets are statistically deficient [38], and therefore, non-availability of an ideal dataset is an issue that needs to be taken into consideration [38].

To develop and evaluate the framework of IDS datasets, few critical characteristics were derived for building a complete and efficient IDS dataset [37]. These characteristics are listed as diversity of attacks, anonymity, available protocols, capturing the complete network traffic, capturing the complete network interaction, defining the complete configuration of the network, feature set, labeled data samples, heterogeneity, and metadata [37]. The CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets are developed keeping these characteristics into consideration. These datasets have used the concept of profiles for generating the datasets in a well-ordered form. Both of these datasets present an in-depth knowledge of attacks conducted and conceptual knowledge about the different application models, network devices, and protocols. The network traffic was captured using the CICFlowMeter that have assigned an appropriate label to the flow and also gives detail about the source and destination address and port number, timestamp, and attack. The simulations of the testing environment consist of network traffic generated from protocols such as HTTP, HTTPS, SSH, and email protocols such as SMTP and POP3.

4.1. Characteristics of the Dataset

The formation of CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets have attracted many researchers to implement different classifiers using these new datasets [39], [40], [41]. The specifications of the datasets are listed in Table 4. The files present in the dataset are used for binary classification as well as multi-class classification. An ideal IDS can be described as the one that is able to detect each attack type precisely, and therefore, to build an efficient IDS the files in the dataset should be merged to have a wide range of attack categories [42]. Moreover, these datasets are developed by taking into consideration the eleven characteristics of an ideal dataset presented in [43] and are listed in Table 5.

4.2. Limitations of Datasets

The observations of CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets have few limitations concerning the data samples and the files created by network flow analysis that can be listed as:

- The data samples generated by the network flow analysis are stored in files and to process these files is a very tedious task as these files have a large number of data instances in each file.

Table 4. Specifications of CIC-IDS-2017 and CSE-CIC-IDS-2018 Datasets [36]

Dataset Name	CIC-IDS-2017	CSE-CIC-IDS-2018
Dataset Type	Multi-class	Multi-class
Year of formation	2017	2018
Duration of Capture	5 Days	10 days
Attack Infrastructure	4 PCs, 1 router, 1 switch	50 PCs
Victim Infrastructure	3 server, 1 firewall, 2 switches, 10 PCs	420 PCs, 30 servers
Features	80	80
Number of Classes	15	18

Table 5. Characteristics for building an ideal Dataset [43]

Characteristic	Description
Network Configuration	It refers to have complete knowledge about the network topology of how the networking devices are connected in the testing environment so that realistic attack scenarios are captured.
Network Traffic	It refers to capturing all the network packets from the host, destination, firewall, and web applications for flow analysis and dataset generation.
Labeled Dataset	It refers to tagging the data instances captured from the network traffic to have a complete understanding of the network interaction.
Network Interaction	It refers to having the complete record of network communication taking place within and outside the network.
Capturing the Traffic	It refers to capturing the functional as well as non-functional network traffic for measuring the DR and FPR of the IDS.
Protocols	An ideal dataset should include all the communication using different protocols whether normal or malicious.
Attacks	The dataset should consist of wide range and updated attacks categories
Anonymity	The dataset should include information from packet header as well as packet payload.
Heterogeneity	The dataset should be collected from varied sources to cover all the details of the procedure carried to detect the attacks.
Features	The dataset should maintain a complete set of well-defined features for classifying the attack.
Metadata	A dataset should have proper documentation describing the testing environment, attack system's infrastructure, victim system's infrastructure, and attack scenarios.

- The files in the dataset can be merged to include each of the attack labels for processing. But combining the instances of each attack type increases the size of the dataset that results in more computing and processing time.
- The dataset also consists of some missing and redundant data records.
- CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets are prone to the issue of high-class imbalance that may result in low accuracy and high FPR of the system [42].

These issues can be handled by preprocessing the data samples, by applying feature engineering or eliminating the missing records. The high-class imbalance can be addressed by relabelling or sampling the data samples and this, in turn, would increase the probability of occurrence of data samples of all classes.

5. Conclusion

The study reviews the datasets developed in the field of Intrusion Detection System (IDS). These datasets have been used for performance evaluation of the ML and DM based IDS. The study revealed that there is a need to update the underlying dataset to identify the recent attacks in the field of IDS with improved performance. This is because the attackers execute attack by using varied processes and technologies. Moreover, the pattern of executing different attacks simulates the need to have datasets with realistic network scenarios. To fulfill the requirement of building an intrusion detection dataset with realistic network traffic and updated network attacks CIC-IDS-2017 and CSE-CIC-IDS-2018 datasets have been introduced. This paper reviews the characteristics of these datasets and also discusses a few shortcomings. In the future, we focus on studying the performance of these datasets with various ML and DM techniques along with incorporating feature engineering and data sampling to address the shortcomings of these datasets.

References

[1] Stevens T. Cyber security and the politics of time. Cambridge University Press; 2016.
[2] Miller NJ, Aliasgari M. Benchmarks for evaluating anomaly-based intrusion detection solutions. California State University, Long Beach; 2018.
[3] Scaife N, Carter H, Traynor P, Butler KR. Cryptolock (and drop it): stopping ransomware attacks on user data. In: 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS). IEEE; 2016. p. 303–312.
[4] Gupta B, Agrawal DP, Yamaguchi S. Handbook of research on modern cryptographic solutions for computer and cyber security. IGI global; 2016.
[5] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E. Anomaly-based network intrusion detection: Techniques, systems and challenges. computers & security. 2009;28(1-2):18–28.
[6] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials. 2016;18(2):1153–1176.
[7] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. computers & security. 2012;31(3):357–374.
[8] Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Computer Science. 2015;60:708–713.
[9] Belavagi MC, Muniyal B. Performance evaluation of supervised machine learning algorithms for intrusion detection. Procedia Computer Science. 2016;89:117–123.
[10] Ektefa M, Memar S, Sidi F, Affendey LS. Intrusion detection using data mining techniques. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). IEEE; 2010. p. 200–203.
[11] Cunningham RK, Lippmann RP, Fried DJ, Garfinkel SL, Graf I, Kendall KR, et al. Evaluating intrusion detection systems without attacking your friends: The 1998 DARPA intrusion detection evaluation. Massachusetts Institute of Technology Lexington Lincoln Lab; 1999.
[12] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE; 2009. p. 1–6.
[13] Dhanabal L, Shantharajah S. A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering. 2015;4(6):446–452.
[14] Tavallaee M, Stakhanova N, Ghorbani AA. Toward credible evaluation of anomaly-based intrusion-detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2010;40(5):516–524.
[15] Deshmukh DH, Ghorpade T, Padiya P. Intrusion detection system by improved preprocessing methods and Naïve Bayes classifier using NSL-KDD 99 Dataset. In: 2014 International Conference on Electronics and Communication Systems (ICECS). IEEE; 2014. p. 1–7.
[16] Mukherjee S, Sharma N. Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology. 2012;4:119–128.
[17] Pervez MS, Farid DM. Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs. In: The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014). IEEE; 2014. p. 1–6.
[18] Farid DM, Harbi N, Rahman MZ. Combining naive bayes and decision tree for adaptive intrusion detection. arXiv preprint arXiv:10054496. 2010;.
[19] Aburomman AA, Reaz MBI. A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security. 2017;65:135–152.
[20] Srivastava S. Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining. International Journal of Computer Applications. 2014;88(10).
[21] Damopoulos D, Menesidou SA, Kambourakis G, Papadaki M, Clarke N, Gritzalis S. Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers. Security and Communication Networks. 2012;5(1):3–14.
[22] Kumar V, Chauhan H, Panwar D. K-means clustering approach to analyze NSL-KDD intrusion detection dataset. International Journal of Soft Computing and Engineering (IJSCE). 2013;.
[23] Ashfaq RAR, Wang XZ, Huang JZ, Abbas H, He YL. Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences. 2017;378:484–497.

[24] Kanakarajan NK, Muniasamy K. Improving the accuracy of intrusion detection using GAR-Forest with feature selection. In: Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015. Springer; 2016. p. 539–547.
[25] Singh R, Kumar H, Singla R. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Systems with Applications. 2015;42(22):8609–8624.
[26] Koch R. Towards next-generation intrusion detection. In: 2011 3rd International Conference on Cyber Conflict. IEEE; 2011. p. 1–18.
[27] Rajahalme J, Conta A, Carpenter B, Deering S. RFC 3697: IPv6 Flow Label Specification. In: The Internet Society; 2004. .
[28] Nehinbe JO. A simple method for improving intrusion detections in corporate networks. In: International Conference on Information Security and Digital Forensics. Springer; 2009. p. 111–122.
[29] Shannon C, Moore D. The caida dataset on the witty worm. Support for the Witty Worm Dataset and the UCSD Network Telescope are provided by Cisco Systems, Limelight Networks, the US Department of Homeland Security, the National Science Foundation. 2004;.
[30] Nechaev B, Allman M, Paxson V, Gurtov A. Lawrence berkeley national laboratory (lbnl)/icsi enterprise tracing project. Berkeley, CA: LBNL/ICSI. 2004;.
[31] Sangster B, O'Connor T, Cook T, Fanelli R, Dean E, Morrell C, et al. Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets. In: CSET; 2009. .
[32] Song J, Takakura H, Okabe Y, Eto M, Inoue D, Nakao K. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security. ACM; 2011. p. 29–36.
[33] Barbosa RRR, Sadre R, Pras A, van de Meent R. Simpleweb/university of twente traffic traces data repository. Centre for Telematics and Information Technology University of Twente, Enschede, Technical Report. 2010;.
[34] Liberatore M, Shenoy P. Umass trace repository. Accessed: May; 2017.
[35] Creech G, Hu J. Generation of a new IDS test dataset: Time to retire the KDD collection. In: 2013 IEEE Wireless Communications and Networking Conference (WCNC). IEEE; 2013. p. 4487–4492.
[36] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In: ICISSP; 2018. p. 108–116.
[37] Sharafaldin I, Gharib A, Lashkari AH, Ghorbani AA. Towards a reliable intrusion detection benchmark dataset. Software Networking. 2018;2018(1):177–200.
[38] Koch R, Golling M, Rodosek GD. Towards comparability of intrusion detection systems: New data sets. In: TERENA Networking Conference. vol. 7; 2014. .
[39] Nicholas L, Ooi SY, Pang YH, Hwang SO, Tan SY. Study of long short-term memory in flow-based network intrusion detection system. Journal of Intelligent & Fuzzy Systems. 2018;(Preprint):1–11.
[40] Radford BJ, Richardson BD. Sequence Aggregation Rules for Anomaly Detection in Computer Network Traffic. arXiv preprint arXiv:180503735. 2018;.
[41] Vijayanand R, Devaraj D, Kannapiran B. Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection. Computers & Security. 2018;77:304–314.
[42] Panigrahi R, Borah S. A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems. International Journal of Engineering & Technology. 2018;7(3.24):479–482.
[43] Gharib A, Sharafaldin I, Lashkari AH, Ghorbani AA. An evaluation framework for intrusion detection dataset. In: 2016 International Conference on Information Science and Security (ICISS). IEEE; 2016. p. 1–6.