A Review on Cyber Security Datasets for Machine Learning Algorithms

Ozlem Yavanoglu
Hacettepe University
Department of Computer Engineering
Ankara, Turkey
milletseverozlem@gmail.com

Murat Aydos
Hacettepe University
Department of Computer Engineering
Ankara,Turkey
maydos@hacettepe.edu.tr

Abstract—**It is an undeniable fact that currently information is a pretty significant presence for all companies or organizations. Therefore protecting its security is crucial and the security models driven by real datasets has become quite important. The operations based on military, government, commercial and civilians are linked to the security and availability of computer systems and network. From this point of security, the network security is a significant issue because the capacity of attacks is unceasingly rising over the years and they turn into be more sophisticated and distributed. The objective of this review is to explain and compare the most commonly used datasets. This paper focuses on the datasets used in artificial intelligent and machine learning techniques, which are the primary tools for analyzing network traffic and detecting abnormalities.**

Keywords— Cyber Security, Data Mining, Artifical Intelligent, Machine Learning, Benchmarking.

I. INTRODUCTION

Cyber security is the set of applying security preventions to provide confidentiality, integrity, and availability of data [1]. Numerous descriptions are made about cyber security in the literature. According to Canongia and Mandarino, "The art of ensuring the existence and continuity of the information society of a nation, guaranteeing and protecting, in Cyberspace, its information, assets and critical infrastructure" [2]. Cyber security is a significant research area because all of the operations based on government, military, commercial, financial and civilians gather, process, and store tremendous volume of data on computers and others [1-3]. In order to be on the defensive side on cyber security, companies require organization of its efforts throughout its whole information system. The components of cyber security consist of network security, application security, mobile security, data security, endpoint security and so on [3].

Over the last few years, the use of the Internet and computer applications has seen an immense expansion and they have turn into the integral part of today's generation of people. With the exponential increase of computer applications and computer networks usage, security is becoming increasingly more significant [4-5]. Attackers are able to potentially use several paths by means of application to do havoc to your business or organization. Figure 1 illustrates some potential attacks and threats to organizations.

All of these paths symbolizes a risk that may or may not be serious enough to warrant attention [4-6]. According to the National Institute of Standards and Technology (NIST), American companies as early as 2017 suffered losses of up to 65.6 billion dollars following IT attacks [6].

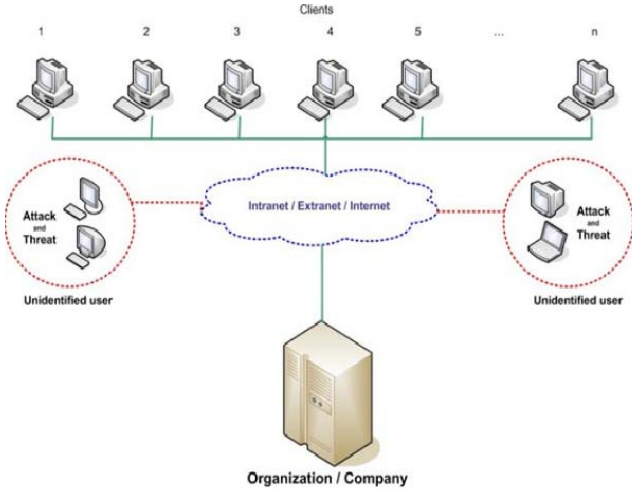

Fig 1. Activities of unidentified as potential attack and threat to organization [2]

The huge increase in the percentage of cyber-attacks has caused artificial intelligent and machine learning based methodologies a crucial part in detecting security threats. In order to provide the best security applications be accepted and appropriate level of security be obtained, security-related benchmarks are very important. From this point of view, they are essential for several types of cyber security research such as intrusion detection system. While there has been a few studies about particular datasets, there has been less about the comprehensive state of security-related datasets. In literature, there are numerous studies in the field of cyber security using various datasets [7-11]. In this study, a comprehensive review of the current publicly available datasets is given. We also provide a general assessment of artificial intelligent and machine learning techniques using these datasets.

The rest of the paper is organized as follows: Part 2 explains the essential security concepts. Part 3 presents summary of the previous studies. Part 3 describes techniques that belong to machine learning algorithm and artificial intelligent. Part 4 discusses major datasets and their characteristic. Finally, part 5 presents observations and concluding remarks.

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

II. Basic Security Concepts

Cyber security is the set of applying security preventions to provide confidentiality, integrity, and availability of data. In this section, we explain the well-known triad of confidentiality, integrity, and availability (CIA) of information security [9-10].

Confidentiality aims to restrict disclosures and to grant access of information to only the authorized people. Thanks to confidentiality, companies are able to protect their sensitive and private assets from unauthorized hands. There are various ways of ensuring confidentiality such as encryption, access controls, and steganography [11].

Integrity requires protecting data in a consistent, precise, and reliable manner. This has to guarantee that data is not altered in the course of a specific period. In order to prevent unauthorized users making modifications, the right processes and actions have to be taken. Hashing, digital signatures, certificates, non-repudiation are the tools and algorithms providing integrity [10-11].

Availability is another security concept that the data and resources should be available when people need to access it, particularly during emergencies or disasters. The cyber security specialists should handle the three common challenges for availability; denial of service (DoS), loss of information system capabilities because of natural disasters and equipment failures during a normal operation [11].

III. RELATED STUDIES

In the literature, there has been considerable amount of studies on the problem of cyber security. There are various widespread approaches in general cyber security solutions. In this section, we have focus on using machine learning and artificial intelligent approaches for cyber security issues.

Chowdhury et al. proposed a new botnet detection method based on topological feature of nodes within a graph. The proposed methodology is able to detect anomaly by searching a limited number of nodes. This methodology is based on self-organizing map (SOM) clustering that belongs to a class of unsupervised system. This study used CTU-13 datasets, the largest dataset that contains bot labeled nodes. Furthermore, this study used another detection algorithm, support vector machine (SVM), for comparison. Experimental results show that proposed methodology could be able to still detect bot with acceptable accuracy by searching few number of nodes [12].

Huseynov et al. proposed a bio inspired computing technique also known as ant colony clustering for detection of botnet attacks. This proposed model is able to explore botnet hosts quickly and precisely while not depending on its traffic payload. At the same time, their approach was tested using two different clustering algorithms that is ATTA-C and K-means for comparison. ISOT dataset was preferred because of its volume [13].

Neethu B. represents a framework that is PCA for feature selection with Naive Bayes in order to develop a network intrusion detection system. In this study, KDDCup 1999 intrusion detection benchmark dataset is preferred for experiments. The results show that the performance of this method achieves higher detection rate, less time consuming and has low cost factor compared to the neural network and tree algorithm based approach. In addition, proposed system provides about 94% accuracy [14].

Rafal and et al. presented a novel method for detecting cyber-attacks targeting web applications. This method was compared with Naive Bayes, AdaBoost, Part and J48, which are machine-learning algorithms. In addition, CSIC 2010 HTTP Dataset is used for assessment of proposed model. This study specifically focused on solutions that are using HTTP protocols to communicate clients with the servers. The authors claimed that this model is able to obtain the higher detection percentage while having lower false positive rate. At the same time, the results show that J48 method is the best approach for this problem and true-positive value is around 0.04 [15].

Nguyen and Franke proposed an adaptive intrusion detection system (A-IDS). This system is able to detect many different types of attacks in the heterogeneous and adversarial network environments. Authors conduct the experiments on two different datasets for benchmarking Web Application Firewalls: the ECML-PKDD 2007 HTTP dataset and the CISIC HTTP 2010. At the same time, Naïve Bayes, Bayes network, decision stump and RBF network, that are machine learning algorithms, are used for comparison with the proposed method. The experimental results illustrated that, in the case of the CSIC 2010 dataset it provides almost 10% and 8% higher accuracies than the best IDS which is the Bayes Network-based IDS, and the Hedge/Boosting algorithm, respectively [16].

Xie and et al. focused on detecting anomalies with a short sequence model. In this study, a novel anomaly detection system is proposed using Support Vector Machine (SVM). ADFA-LD is used for conducting experiments. For this experiments, k values were selected k = 3, 5, 8, 10 and the best achievement is obtained with k=5, where average ACC of 70% is achieved at a FPR of around 20%. The experimental result represents that it not only provide a satisfactory achievement, but also decrease the computational cost largely [17].

Zamani and Movahedi represent several models for detecting intrusion. In this study, these models are divided based on classical artificial intelligence (AI) and based on computational intelligence (CI) such as genetic algorithms and fuzzy logic. They conducted various experiments and compared their algorithms' performance. Experimental results shows that decision tree algorithm has achieved the best results. On the other hand, this study explained how different features of CI models could be used to build effective IDS [18].

In order to efficiently detect various types of network intrusions, Hoque1 et al. proposed an intrusion detection system (IDS) based on genetic algorithms. In this study, parameters and evolution processes of GA were explained in details. Proposed model used evolution theory for information evolution in order to filter the traffic data and thus decrease the complexity. In addition, KDD99 benchmark dataset used in order to evaluate the performance of the model. The experimental results show that this model has achieved reasonable detection rate [19].

Wang and Paschalidis proposed a novel approach that has two stage in order to detect the presence of a botnet and to identify the bots. First stage is relevant to becoming aware of anomalies by leveraging large deviations of an empirical distribution. In addition, this stage suggests two techniques for creating the empirical distribution. First technique is a flow-based approach estimating the histogram of quantized flows and latter is a graph based approach estimating the degree distribution of node interaction graphs. In order to detect the bots, second stage uses social network community in a graph that captures correlations of interactions among nodes over time. For the experiments, they used real-world botnet traffic that is CTU-13 dataset [20].

Bhuyan et al. introduced a new approach to create unbiased full feature real-life network intrusion datasets in order to compensate for the crucial lack of the available datasets. They created a significant amount of an intrusion dataset in the development and validation operation of detection systems. In addition, this study explains a set of requirements for creating an efficient dataset. Finally, six different attack scenarios were created and discussed in this study [21].

Wijesinghe et al. focus on detecting a range of botnet families by analyzing network traffic flows. Their proposed method consists of two parts. First parts is that they define appropriate dataset templates with more relevant features in order to detect botnet from IP flows. Second part used IP flow data for detecting botnet behaviors in unlabeled traffic. In this study, they used public available IPFIX dataset. This approach is a new methodology and it contributed to available IP flow based botnet detection studies [22].

Haddadi et al. have analyzed various botnet detection approaches based on the model used and type of data employed. BotHunter and Snort based on public rule based systems are two of approaches. Other approaches are based on data mining techniques like packet payload based and traffic flow based. This study makes use of five publicly available botnet data sets such as CAIDA, ISOT, etc. They conducted several experiments using C4.5, Knn (k-nearest neighbors), SVM, Bayesian Networks. Experimental results show that the performance of the flow based system is higher or similar to the results reported in the literature [23].

Bhuyana et al. proposed an empirical study using different information metrics in order to handle important security problems such as detection of both low-rate and high-rate DDoS attacks. They conducted several experiments using four significant information entropy measures: Hartley entropy, Shannon entropy, Renyin++s entropy and Renyin++s generalized entropy for detecting DDoS attacks of various types. CAIDA and TUIDS DDoS datasets are used for showing efficiency and effectiveness of each metric for DDoS detection [24].

Hoque et al. proposed a novel statistical methodology in order to analyze DDoS attack from normal traffic. This methodology called as Feature Feature Score (FFSc). This study used three features from network traffic. These features are entropy of source IPs, variation of source IPs and packet rate. The success of the proposed model is evaluated with CAIDA DDoS 2007 and MIT DARPA datasets. The experimental results show that proposed model yields 98% detection accuracy on the normalized CAIDA dataset [25].

Kato and Klyuev have developed an DDoS attack detection system. Also, this study analyzed the characteristics of DDoS attacks. This system used SVM with an RBF (Gaussian) kernel from machine learning. To compare the performance of the proposed system, three types of training and test datasets including different patterns and different number of patterns were created. For evaluating success of system, precision, recall, negative predictive value (NPV) were calculated. Development system has achieved successful results with more than 85% accuracy with all types of dataset [26].

Saad et al. focused on detecting P2P bots that represents the newest and most challenging types of botnets currently available. In order to detect P2P botnet command and control (C&C) phase, they proposed the characterization of network traffic behaviors.

In this study, they used five machine-learning algorithms that are nearest neighbors, linear support vector machine, artificial neural network, and naïve bayes. The experimental results show that true detection rate of the P2P Botnet C&C is above 90% for the Support Vector Machine, Artificial Neural Network and the Nearest Neighbors Classifier and the total error rate is less than 7% [27].

A number of related detection systems are compared and the results shown in Table 1. In particular, we compare the machine learning techniques used for developing the detection systems datasets used for experiments, evaluation methods considered, baseline classifiers for comparisons, etc. in relevant studies.

IV. TECHNIQUES

A. Machine Learning

According to Stanford computer science professor Andrew Ng, Machine learning (ML) is "the science of getting computers to act without being explicitly programmed." [32]. The primary aim of ML is to build models that can take input data and utilize statistical analysis in order to forecast an output value within an suitable range. In the field of computer science, ML is one of the fastest expanding areas with comprehensive applications. ML algorithms are often classificatied as supervised, unsupervised and Reinforcement Learning. Supervised algorithms are the most commonly used in the machine learning algorithms. In addition, supervised algorithms can be further grouped into regression and classification. In literature, several machine learning algorithms used [10-40]. Commonly used machine-learning algorithms are;

- Linear Regression
- Logistic Regression
- Decision Tree
- SVM
- Naive Bayes
- KNN
- K-Means
- Random Forest

B. *Artifical Intelligence*

AI is a field of scientific research to increase computing power, to develop productive algorithms and well organized knowledge. AI applies for solving complicated problems that cannot be solved without combining intelligence, discovering the hidden patterns from data and developing intelligent machines [18].

AI has numerous applications on knowledge representation, information retrieval, speech recognition, understanding natural language, computer vision, bioinformatics, expert systems, robotics, game playing, and cyber defense with the help of various algorithms like artificial neural network ,genetic algorithms, artificial immune systems, particle-swarm intelligence, stochastic algorithms, and fuzzy logic [19, 20].

Artificial Neural Networks (ANNs), which is a technique of AI, are set of computer algorithms that are biologically inspired to simulate the way in which the human brain neuron processes information [40]. ANNs gather their knowledge by detecting the patterns and relationships among data and learn through their architectures, transfer functions and learning algorithms [40].

There are many types of neural networks for various applications available in the literature [39]. Multilayered perceptron (MLP) type neural networks are the simplest and most commonly used neural network architectures [40].
MLPs are trained with many learning algorithms. Levenberg-Marquardt (LM) is one of most preferred training algorithms for MLPs.

Table 1. Outline of the Studies Presented in the Literature

Study	Technique	Dataset	Problem Domain	Evaluation Method	Feature Selection
[25]	Statistical Method	CAIDA DDoS 2007. MIT DARPA datasets	DDoS attack detection	Accuracy	Yes
[23]	C4.5, SVM, KNN Bayesian Networks	Zeus (Snort), Zeus (NETRESEC), Zeus-2 (NIMS), Conficker (CAIDA) and ISOT-Uvic	Botnet detection	Detection Rate, False Positive Rate	Yes
[12]	SOM	CTU-13	DDoS attack detection	Accuracy	Yes
[14]	Naïve Bayes, PCA algorithm	KDDCup 1999	Intrusion Detection	False Positive Rate	Yes
[26]	SVM	CAIDA DDoS 2007	DDoS attack detection	Precision, Recall, Negative Predictive Value	Yes
[15]	Naive Bayes, AdaBoost, Part and J48	CSIC 2010 HTTP Dataset	Web Applications Attack	False Positive Rate	No
[16]	Naïve bayes, bayes network, decision stump RBF network	ECML-PKDD 2007 HTTP, CSIC HTTP 2010	Web Applications Attack	False Positive Rate	No
[17]	k nearest neighbour (kNN)	ADFA Linux data	Host-based Anomaly Detection	Accuracy	No
[19]	Genetic Algorithm	KDDCup 1999	Intrusion Detection	Detection rate (DR)	Yes
[24]	Information Metrics	KDD Cup 1999, CAIDA , TUIDS DDoS	DDoS Attack Detection	N/A	No
[27]	NNC ANN SVM NBC GBC	ISOT	Botnet Detection	true detection rate, Error Rate	Yes
[28]	SVM,J48, Naive Bayes, Logistic Regression	ISOT, UNSW-NB-15	Cloud Security	True Positive , False Negative	No
[29]	Decision Trees Language Modeling TF-Based	ECML-PKDD 2007 Dataset	HTTP Attacks	precision, recall	Yes
[30]	KNN-SVM	KDD99	DDoS attack detection	True Positive Rate, False Positive Rate	Yes
[31]	Adaptive Neuro-Fuzzy Inference System	KDD99, CAIDA	DDoS attack detection	Accuracy	No
[32]	Generic-Feature-Selection (GeFS)	CSIC 2010 HTTP Dataset	Feauture Selection	Accuracy	Yes
[33]	Random Forest	KDD99	Feauture Selection	Accuracy	Yes
[34]	RBF, SVM	KDD99	Network Intrusion Detection	True Positive , False Negative	Yes
[35]	Adaptive Time Dependent Transporter Ants Clustering	ISOT	Botnet Detection	Accuracy	No

V. CYBER SECURITY DATASETS

Nowadays, several research groups put together many type of data both for their own study purposes and to provide data to community repositories. This section explains the existing security-related datasets using machine learning and artificial intelligent research.

A. KDD Cup 1999 Dataset (DARPA1998)

DARPA 1998 has gathered and deal out the first standard data by MIT Lincoln Laboratory under Defence Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL) sponsorship to evaluate computer network intrusion detection systems. KDD Cup 1999 is part of the data collected from MIT Lincoln Labs, includes tcpdump and BSM list files. This dataset is based on the data captured in DARPA'98 IDS evaluation program and prepared by Stolfo et al. [5]. Also, this dataset is considered benchmark data for assessment of intrusion detection systems. The data includes four main categories of attacks that are Denial-of-Service (DoS), user-to-root (U2R), Remote to Local Attack (R2L) and Probing Attack. Also, there are three content features and thirty-eight numerical features in the dataset. The features consist of basic features of individual TCP connections, content features within a connection suggested by domain knowledge and traffic features computed using a two-second time window. KDD'99 is one of the most wildly popular used data set to evaluate performance of anomaly detection methods. As of today, there are thirty researches using KDD dataset [12-17].

B. ECML-PKDD 2007 Dataset

The ECML-PKDD 2007 dataset was created for the European Conference on Machine Learning and Knowledge Discovery in 2007. The ECML/PKDD Discovery Challenge was a data mining competition held in conjunction with the 18th European Conference on Machine Learning (ECML). Table II shows characteristics of ECML/PKDD 2007.

Table I. Features of ECML/PKDD Dataset

	Training Set	Test Set
Total Request	50.116	70,143
Valid Request	35,006 (70%)	42,006 (60%)
Attacks	15,110 (30%)	28,137 (40%)
Cross Site Scripting	12%	11%
SQL Injection	17%	18%
LDAP Injection	15%	16%
XPATH Injection	15%	16%
Path traversal	20%	18%
Command Execution	23%	23%
SSI	13%	12%

The dataset is described in extensible markup language (XML). All of the sample is represented by a unique id and consists of the three main parts that are context, class and query [18-25].

Context parts include following features:

- Operating system running on the web server, HTTP Server targeted by the request, Is the XPATH technology understood by the server, Is there an LDAP database on the Web Server?, Is there an SQL database on the Web Server?

Query parts include features that are method, protocol, uri, query, headers and body.

C. ISOT (Information Security and Object Technology) Dataset

ISOT (Information Security and Object Technology) dataset is a combination of openly available various botnets and normal datasets that contains 1,675,424 total traffic flow. For malicious traffic in ISOT, it was collected from French chapter of honeynet project that consist of Storm and Waledac botnets. Non-malicious traffic was obtained from Traffic Lab Ericson Research in Hungary. After that, this traffic was combined with another dataset that is created by Lawrence Berkeley National Lab (LBNL). This compilation contains general traffic from numerous type of applications besides that HTTP web browsing, World of Warcraft traffic, and traffic from Azureus bittorent client. Thus, this traffic is considerable big dataset for Ericson Lab. LBNL network trace covered 22 subnets from 2004 to 2005. Moreover, LNBL traffic consists of a medium-sized enterprise network and involves five huge datasets [28].

D. HTTP CSIC 2010 Dataset

The HTTP CSIC 2010 dataset involves several thousands of web requests that generated automatically and developed at Information Security Institute of CSIC (Spanish Research National Council). The dataset can be used for testing web attack protection systems. This data consist of 6,000 normal requests and more than 25,000 anomalous requests and HTTP requests are labeled as normal or anomalous. For convenience, the dataset are split into three different subsets that are training, anomalous and testing. The anomalous requests refer to a comprehensive field of application layer attacks. In this dataset, there are three types of attacks that are static, dynamic and unintentional illegal requests. For example, SQL injection, CRLF injection, cross-site scripting, buffer overflows, etc are dynamic attacks. Static attacks try to request hidden resources. These requests include obsolete files, session ID in URL rewrite, configuration files, default files, etc. Unintentional illegal requests do not have malicious intention, however they do not follow the normal behavior of the web application and do not have the same structure as normal parameter values (for example, a telephone number composed of letters). This dataset has been successfully used for web detection in previous works [40-46].

E. CTU-13 (Czech Technical University) Dataset

CTU-13 (Czech Technical University) dataset is the combination of seizures of 13 different malware in a nonfictional network environment. The aim of this dataset is to capture real mixed botnet traffic. Infected hosts generated botnet traffic and verified normal hosts generated normal traffic. Lastly, Background traffic is a remainder of the traffic that we do not know what it is for sure. The CTU-13 dataset includes thirteen captures of different botnet samples, also

known as scenarios. Each of all scenarios was executed with a particular malware that used various protocols and carried out several actions. This dataset is one of the largest and more labeled into existing datasets and created by CTU University of Prague in Czech Republic in 2011. Firstly, Grill et al. have used the CTU-13 dataset. This study compared various botnet detection methodologies using CTU-13 dataset and proposed a novel error metric [14]. In this study, to evaluate performance of botnet detection, BClus and The Cooperative Adaptive Mechanism for Network Protection (CAMNEP) and BotHunter algorithms were used. This dataset has been used in lots of studies. In 2014, Grill et al. used this data set to measure results of local adaptive multivariate smoothing (LAMS) model on the NetFlow anomaly detection. False alarm rate of anomaly detection on intrusion detection systems has been able to be reduced thanks to proposed model [16]. The details of the scenario is shown in Table I with properties. The advantage of using this dataset is that it is carefully labeled dataset and capturing process conducted in controlled environment [25-30].

Table II. Amount of data on each botnet scenario

Dataset	Duration (h)	NetFlows	Size (GB)	Bot name	Number of bots	Botnet flow
1	6.15	2,824,637	52	Neris	1	39933 (1.41%)
2	4.21	1,808,123	60	Neris	1	18839 (1.04%)
3	66.85	4,710,639	121	Rbot	1	26759 (0.56%)
4	4.21	1,121,077	53	Rbot	1	1719 (0.15%)
5	11.63	129,833	37.6	Virut	1	695 (0.53%)
6	2.18	558,920	30	Menti	1	4431 (0.79%)
7	0.38	114,078	5.8	Sogou	1	37 (0.03%)
8	19.5	2,954,231	123	Murlo	1	5052 (0.17%)
9	5.18	2,753,885	94	Neris	10	179880 (6.5%)
10	4.75	1,309,792	73	Rbot	10	106315 (8.11%)
11	0.26	107,252	5.2	Rbot	3	8161 (7.6%)
12	1.21	325,472	8.3	NSIS.ay	3	2143 (0.65%)
13	16.36	1,925,150	34	Virut	1	38791 (2.01%)

F. The ADFA Datasets

In the field of host-based anomaly detection, most of the existing benchmark data sets, such as UMN [2] and DARPA [3] intrusion detection data sets, were compiled a decade ago and have failed to reflect the characteristics of modern computer systems. In 2013, Australian Defence Force Academy Linux Dataset has been released by the Australian Defence Force Academy in University of New South Wale. In order to evaluate host based intrusion detection system, ADFA dataset (Linux dataset) was generated on a Ubuntu Linux 11.04 host OS with Apache 2.2.17 running PHP 5.3.5. FTP, SSH, MySQL 14.14, and TikiWiki were started. This dataset involves normal and attack Linux based system calls traces. When a sampling stage, the host that is configured to represent a modern Linux server captures the system call traces where legitimate programs are operated as usual. Subsequently, the cyber-attacks, i.e., Hydra-FTP, HydraSSH, Adduser, Java-Meterpreter, Meter-preter and Webshell, are launched in turn against the host, each of which results in 8-20 abnormal traces. Table III. has shown the composition of ADFD-LD

Table III. The composition of ADFD-LD

Trace Type	Number	Label
Training	833	normal
Validation	4373	normal
Hydra-FTP	162	attack
Hydra-SSH	148	attack
Adduser	91	attack
Java-Meterpreter	125	attack
Meterpreter	75	attack
Webshell	118	attack

The aim of ADFA dataset is to take the place of existing benchmark data sets, because these benchmark datasets have failed to reflect the characteristics of modern computer systems.

G. UNSW-NB15 Dataset

UNSW-NB 15 data set was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS). This dataset contains approximately one hour of anonymized traffic traces from a DDoS attack in 2007 [35-39].

This dataset represent nine types of major attacks that are Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. In order to classify this dataset, IXIA PerfectStorm tool has achieved report from the attack data. Table IV illustrates types of modern attack in this dataset.

Table IV. Features of UNSW-NB15 Dataset

Category	**Traning set**	**Testing set**
Normal	56.000	37000
Analysis	2.000	677
Backdoor	1.746	583
DoS	12.264	4089
Exploits	33.393	11.132
Fuzzers	18.184	6.062
Generic	40.000	18.871
Reconnaissance	10.491	3.496
Shellcode	1.133	378
Worms	130	44
Total Records	175.341	82.332

There are 49 features in this dataset. In order to extract features, Argus, Bro-IDS tools were used and 12 models were developed. Features are categorized into only five groups that are flow features, basic features, content features, time features and additional generated features. Compared to existing dataset, this dataset has several attack families that ultimately reflect modern low foot print attacks [40].

VI. CONCLUSION

The protection of computer systems from cyber-attacks is one of the main issues for national and international security. Various researches have been conducted using several datasets and artificial intelligent and machine learning play a significant role in protection of computer systems .In this paper, we have outlined a comprehensive classes of various datasets along with their advantages and disadvantages. In the future, we are going to plan generating a new dataset and make it publicly available.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their constructive comments and valuable suggestions.

REFERENCES

[1] Sumeet, D., Xian, D., “Data Mining and Machine Learning in Cybersecurity”. CRC press, 2016.

[2] Claudia, C., Mandarino R. Jr., "Cybersecurity: The New Challenge of the." Handbook of Research on Business Social Networking: Organizational, Managerial, and Technological Dimensions: Organizational, Managerial, and Technological Dimensions, 2011.

[3] Paul, T.,” Cyber Security Threats”, 2010.

[4] Rossouw, V., Niekerk, J. V.. "From information security to cyber security." computers & security 38, 2013.

[5] Fraley, J. B., Cannady, J., "The promise of machine learning in cybersecurity" SoutheastCon, 2017. IEEE, 2017.

[6] https://www.symantec.com/content/dam/symantec/docs/other-resources/web-application-firewall-owasp-top-10-2017-coverage-en.pdf (2017) accessed

[7] Buczak, A. L., Guven, E., "A survey of data mining and machine learning methods for cyber security intrusion detection." IEEE Communications Surveys & Tutorials 18.2 , 2016.

[8] Thuraisingham, B., "Data mining for security applications." Embedded and Ubiquitous Computing, 2008. EUC'08. IEEE/IFIP International Conference on. Vol. 2. IEEE, 2008.

[9] Meshram, A., Haas, C., "Anomaly detection in industrial networks using machine learning: a roadmap", Machine Learning for Cyber Physical Systems. Springer Berlin Heidelberg, 2017.

[10] Feily, M., Alireza S., Sureswaran R., "A survey of botnet and botnet detection", Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on. IEEE, 2009.

[11] Salem, M.B., Shlomo H., Salvatore J. S., "A survey of insider attack detection research", Insider Attack and Cyber Security, 2008.

[12] Chowdhury, S., "Botnet detection using graph-based feature clustering." Journal of Big Data 4.1, 2017

[13] Neethu, B., “Adaptive Intrusion Detection Using Machine Learning”, International Journal of Computer Science and Network Security (IJCSNS), 13(3), 118, 2013.

[14] Kozik, R., Choraś, M., Renk, R., Hołubowicz, W., “A Proposal of Algorithm for Web Applications Cyber Attack Detection”, In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 680-687). Springer, Berlin, Heidelberg

[15] Nguyen, H. T., FRANKE, K., “Adaptive Intrusion Detection System via online machine learning”, In: Hybrid Intelligent Systems (HIS),12th International Conference on. IEEE, 2012.

[16] Xie, M., Jiankun, H., Jill, S., "Evaluating host-based anomaly detection systems: Application of the one-class svm algorithm to adfa-ld", Fuzzy Systems and Knowledge Discovery (FSKD), 11th International Conference on. IEEE, 2014.

[17] Zamani, M., Mahnush M., "Machine learning techniques for intrusion detection”, arXiv preprint arXiv:1312.2177, 2013.

[18] Hoque, M. S., Mukit, M., Bikas, M., Naser, A., “An implementation of intrusion detection system using genetic algorithm”, arXiv preprint arXiv:1204.1336, 2012.

[19] Wang, J., Ioannis Ch Paschalidis, "Botnet detection based on anomaly and community detection", IEEE Transactions on Control of Network Systems 4.2 2017.

[20] Bhuyan, M. H., Dhruba K. B., Jugal K. K., "Towards Generating Real-life Datasets for Network Intrusion Detection", IJ Network Security 17.6, 2015.

[21] Wijesinghe, U., Udaya T.,Vijay V., "An enhanced model for network flow based botnet detection", Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015). Vol. 27. 2015.

[22] Haddadi, F., Le L., D., Porter, Zincir-Heywood L., “On the Effectiveness of Different Botnet Detection Approaches”, In ISPEC (pp. 121-135), 2015.

[23] Bhuyan, M. H., Bhattacharyya, D. K., Jugal K. Kalita, "An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection", Pattern Recognition Letters, 2015.

[24] Hoque, N,, Dhruba K. B,, Jugal K. Kalita, "A novel measure for low-rate and high-rate DDoS attack detection using multivariate data analysis", Communication Systems and Networks (COMSNETS), 8th International Conference on. IEEE, 2016.

[25] Kato, K., Vitaly K., "An Intelligent DDoS Attack Detection System Using Packet Analysis and Support Vector Machine", IJICR,478-485, 2014.

[26] Sherif, S., Issa, T., Ali A., Ghorbani, B. S., David, Z., Wei Lu, John Felix, Payman Hakimian, "Detecting P2P botnets through network behavior analysis and machine learning", Proceedings of 9th Annual Conference on Privacy, Security and Trust (PST2011), Montreal, Quebec, Canada, 2011.

[27] Bhamare, D., Salman, T., Samaka, M., Erbad, A., Jain, R., “Feasibility of Supervised Machine Learning for Cloud Security”, In Information Science and Security (ICISS), 2016 International Conference on (pp. 1-5), 2016,

[28] Galllagher, B., Eliassırad, T., “Classification of http attacks: a study on the ECML/PKDD 2007 discovery challenge”, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2009.

[29] Ahmad, Y.U., Nur, I., Ali, S., “An Evaluation on KNN-SVM Algorithm for Detection and Prediction of DDoS Attack”, In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer International Publishing, 2016.

[30] Kumar, K., Arun Raj, P., Selvakumar, S., "Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems", Computer Communications, 2013.

[31] Torrano-Gimenez, C., Nguyen, H., Álvarez, G., Petrovic, S., Franke, K., “Applying Feature Selection to Payload-Based Web Application Firewalls”, In Proc. of International Workshop on Security and Communication Networks (IWSCN 11), pp. 75-81. Editor Patric Bours. Gjøvic (Noruega). ISBN: 978-82-91313-67-2. 18-20 Mayo, 2011.

[32] Hasan, M., Nasser, M., Ahmad, S., Molla, K., “Feature Selection for Intrusion Detection Using Random Forest”, Journal of Information Security, 2007.

[33] Mrutyunjaya, P., Abraham, A., Ranjan Patra, M., "A hybrid intelligent approach for network intrusion detection." Procedia Engineering 30, 2012.

[34] Lippmann, R. P., Fried, D. J., Graf I., J. W., Haines, K., D. McClung, D., Webber, S., Wyschograd, D., Cunninghan, R., Zissman, M, “Evaluating Intrusion Detection Systems: The 1998 DARPA off-line intrusion detection evaluation”, In Proc. of DARPA Information Survivability Conference and Exposition (DISCEX00), Hilton Head, South Carolina, January 25-27. IEEE Computer Society Press, Los Alamitos, CA, 1226, 2000.

[35] Lippmann, R., Haines, J. W., Fried, D. J., Korba J.K, “The 1999 DARPA Off-Line Intrusion Detection Evaluation”, In Proc. Recent Advances in Intrusion Detection (RAID2000). H. Debar, L. Me, and S. F. Wu, Eds. Springer-Verlag, New York, NY, 162182, 2000.

[36] McHugh, J., “Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory”, In Proc. of ACM Transactions on Information and System Security (TISSEC) 3(4), pp. 262-294, 2000.

[37] PerezVillegas, A., TorranoGimenez, C., Alvarez G., “Applying Markov Chains to Web IntrusionDetection”, In Proc. of Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2010), pp. 361-366. Publicaciones urv. Tarragona (España), 7-10 Septiembre 2010.

[38] TorranoGimenez C., PerezVillegas,A., Alvarez, G., “An anomaly-based approach for intrusion detection in web traffic”, Journal of Information Assurance and Security, vol. 5, issue 4, pp. 446-454. ISSN 1554-1010, 2010.

[39] TorranoGimenez, C., Perez-Villegas, A., Alvarez, G., “A Self-Learning Anomaly-Based Web Application Firewall”, In Proc. of 2nd International Workshop in Computational Intelligence in Security for Information Systems (CISIS 09). Advances in Intelligent and Soft Computing, vol. 63, pp. 85-92, Springer-Verlag. A. Herrero, P. Gastaldo, R. Zunino, E. Corchado, editores. Burgos (España), 23-26 Septiembre, 2009.

[40] Torrano-Gimenez, C., Perez-Villegas, A., Alvarez, G., “An Anomaly-based Web Application Firewall”, In Proc. of International Conference on Security and Cryptography (SECRYPT 2009), pp. 23-28. INSTICC Press. E. Fernández-Medina, M. Malek, J. Hernando, editores. Milán (Italia), 7-10 Julio, 2009.