Review Article

A Detailed Analysis of Benchmark and Recent Datasets for Network Intrusion Detection System

Comment [I1]: Formatting, Check alignments / Okay or not

ABSTRACT

The enormous increase in the use of the Internet in daily life provided an opportunity for the intruder attempt to compromise the security principles of confidentiality, availability, and integrity. As a result, organizations are working to increase the level of security by using attack detection techniques such as the Network Intrusion Detection System (NIDS), which monitors and analyzes network flow and attacks detection. There are a lot of researches proposed to develop the NIDS and depend on the dataset for the evaluation. Datasets allow evaluating the ability in detecting intrusion behavior. This paper introduces a detailed analysis of benchmark and recent datasets for NIDS. Specifically, we describe eight well-known datasets that include: KDD99, NSL-KDD, KYOTO 2006+, ISCX2012, UNSW-NB 15, CIDDS-001, CICIDS2017, and CSE-CIC-IDS2018. For each dataset, we provide a detailed analysis of its instances, features, classes, and the nature of the features. The main objective of this paper is to offer overviews of the datasets are available for the NIDS and what each dataset is comprised of. Furthermore, some recommendations were made to use network-based datasets.

Keywords:KDD99; NSL-KDD; KYOTO 2006+; ISCX2012; UNSW-NB 15; CIDDS-001; CICIDS2017; CSE-CIC-IDS2018.

1. INTRODUCTION

Network security has become increasingly important with the rising growth of computer networks and the increasing use of computer applications on these networks. The big challenge facing network engineers and researchers today is to identify malicious activities in a host or over a network [1]. The Cyber security research area looks at the ability to act proactively to mitigate or prevent attacks.

Network Intrusion Detection System (NIDS) is placed at a strategic point in the network where it monitors all the traffic, it analysis the traffic to detect possible attacks. Mostly, NIDS follows one of the two major detection methods: anomaly-based intrusion detection system (AIDS) and signature-based intrusion detection system. In addition, a lot of researchers have proposed hybrid methods. Signature-based detection is quite popular in commercial applications for designing effective commercial NIDS, it is designed to detect known attacks that are preloaded in the NIDS datasets. Anomaly-based detection is limited to academics for research and development, it compares current user activities against

Comment [I2]: A - Capital letter (Anomaly)

Comment [I3]: No need another place not using also signature-based intrusion detection system (SIDS) it is not a mention

predefined profiles is used to detect abnormal behaviors that might be intrusions. Anomaly-based detection methods are prime in detecting network-level attacks, it is an effective way to detect unknown attacks [2-5]. Anomaly-based detection methods are better than Signature-based detection methods in the detection of new attacks [6, 7]. A hybrid detection is combined two methods to overcome disadvantages in signature-based detection and obtain advantages for anomaly-based detection [8]. But in general, NIDS needs existing information to detect future attacks.

Datasets are needed to train and evaluate anomaly-based network intrusion detection systems, given a labeled dataset in which each data instance is assigned to the class normal or attack, the number of detected attacks may be used as evaluation criteria [9]. Moreover, benchmark datasets are a good basis for evaluating and comparing the quality of different NIDS, which researchers in the field can use to train and test their models [10, 11]. The machine learning model is trained to distinguish between normal and malicious activity [12].

Various datasets have appeared since 1998 until now, some of these datasets suffer from providing volume and variety of network traffic, and others do not have different or new attack patterns, while others lack metadata information. Many researchers have used various machine and deep learning techniques depending on the presence or absence of labeled datasets. This paper concentrates on machine learning techniques, both supervised and unsupervised learning methods that are used by researchers in this field to detect attacks in the network. The main objective of this paper is to provide researchers idea about what the benchmark datasets are publicly available for the evaluate NIDS and what each dataset is comprised of in terms of instances, features, classes, and the nature of the features.

The rest of this paper is organized as follows: Section 2 offers the related work of NIDS. Section 3 provides a detailed analysis of various benchmark datasets. Section 4 offers discussions and recommendations for the use of network datasets. Finally, this paper concludes with future work.

2. RELATED WORK

A lot of researches focus on analyzing benchmark datasets. Almost NIDS researches often focus on analyzing a single dataset of NIDS evaluation or introduce a general review of datasets, a little of researches that presented a detailed analysis of benchmark and recent datasets for NIDS such as:

Panigrahi et al.[3] introduced a detailed analysis of the most recent dataset namely the CICIDS2017 dataset, it consisting of the latest attacks and features. This dataset draws the interest of many researchers because it represents attacks that old datasets did not address. Various lack of the dataset have been studied and outlined. The presented a detailed characteristics of the CICIDS2017 dataset only.

Khraisat et al.[4] demonstrated a survey of NIDS approaches, types, and technologies with their advantages and limitations. The various machine learning techniques that are suggested to detect zero-day attacks are displayed. However, such approaches may have the problem of generating and updating the information about new attacks and poor accuracy or generate high false alarms. Summarized recent studies and explored contemporary models for improving performance NIDS as a solution to overcome on NIDS problems. Additionally, the most common public datasets used in NIDS were showed.

Ring et al.[9] presented a survey about the datasets used for NIDS and describe the underlying packet and flow-based network data in detail. The paper identified fifteen different properties to evaluate the suitability of individual datasets for specific evaluation

Comment [I4]: Why mention need to be labeled? this only for the supervised method
There is an unsupervised method no using labeled. Should be explained when using methods whats procedure need

Comment [I5]: Why mention binary class why not mention multi-class,
Explain for a technical data instance assigned

Comment [I6]: Explain how can using Data set and separate data to use for train and test, There are models/methods like
-K-fold (Cross Viladitaion)
-Synthetic Minority Oversampling TEchnique (SMOTE)

Comment [I7]: Can mention about related work, state of the art, more clear like highlighted one statement, for gaps, something issues,...
And add one paragraph about your work will present ...

scenarios, it also highlighted the peculiarities of each dataset. Furthermore, they provided a discussion and observations and also provided some recommendations for the use and the creation of NIDS datasets.

Hamid et al. [11] provided a review for six benchmark (DARPA98, KDD99, NSL-KDD, Caida DDoS, UNM, and UNSW-NB15) datasets. Moreover, they introduced a detailed discussion for three datasets (KDD99, NSL-KDD, and UNSW-NB 15) based on the number of instances, features, classes, and nature of features. In experimental, they used the K-NN classifier on these six datasets and demonstrated the K-NN classifier algorithm performed better on the NSL-KDD dataset and achieved high performance.

The study by Ferrag et al.[13] showed a survey of deep learning approaches for cybersecurity intrusion detection. They showed thirty-five popular cyber datasets and presented a classification of these datasets into seven categories. Furthermore, seven deep learning models were also analyzed. They used deep learning approaches on two recent (CSE-CIC- IDS2018 and Bot-IoT) datasets and compared performance based on false alarm rate, accuracy, and detection rate. This study introduced a general review on datasets that used for NIDS.

Hindy et al. [14] indicated to specify research gaps, and lack of existing datasets and their effect on the building NIDS, and the growing number of complex attacks. It also provided researchers with two basic pieces of information; a review of well-known datasets, and analyze their use and their effect on the evolution of NIDS.Furthermore, the paper showed that only 33.3% of the attacks were covered by current NIDS research. Additionally, current datasets demonstrated a clear shortage of real network attacks, attack representation, which together border the detection accuracy of attacks for NIDS.

3. INTRUSION DETECTION DATASETS

Comment [I8]: Should focus the address of this section only NIDS, when said IDS their Host (HIDS) so better indicator Network (NIDS), this you're looking only.

Can add more details for challenges of NIDS about the dataset.
And can make a list of kinds of intrusions to networks, like definition / declarative, of type NIDS

Datasets play an important role in evaluating NIDS, which can be used for experiments and validating new techniques [15]. Researchers relied on benchmark datasets to evaluate their results. However, currently available datasets lack realistic characteristics of recent network traffic [16]. Moreover, NIDS is unable to adapt to constant changes in networks. Networks are constantly changing, for this reason depending solely on old datasets does not help the progress of NIDS. The process of generating new datasets should consider this constant change fact in the network [14]. The detailed analysis of the datasets illustrate in the following subsections.

3.1 KDD99 Dataset

The KDD99 was created by MIT and utilized in the International Knowledge Discovery and Data Mining Tool Competition [17]. The benchmark dataset for IDS was KDD99 released by DARPA [15]. The dataset was prepared in 1999 and has become the most widely used dataset for the evaluation of anomaly detection although KDD99 dataset is more than 20 years old [18].KDD99 dataset consists of 4,898,431 instances each of which consists of 42 features. Table 1 shows KDD99 dataset features.

Table 1. KDD99 dataset features.

No	Features	No	Features
1	duration lenght	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count

4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	lnum_compromised	34	dst_host_same_srv_rate
14	lroot_shell	35	dst_host_diffsrv_rate
15	lsu_attempted	36	dst_host_same_src_port_rate
16	lnum_root	37	dst_host_srv_diff_host_rate
17	lnum_file_creations	38	dst_host_serror_rate
18	lnum_shells	39	dst_host_srv_serror_rate
19	lnum_access_files	40	dst_host_rerror_rate
20	lnum_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_hot_login	42	Class

KDD99 contents a total of 22 training attacks types and one normal, with 17 additional types in the testing data only [19]. The 41 features labeled as either special attack type (DOS, U2R, R2L, and Probe) or normal. It is believed that attacks can be detected with the knowledge learned from the registered attacks [20]. Although widely used, this dataset has inherent flaws [2]. Attack types in KDD99 dataset can be fall in one of the main four categories:

1. **Denial of Service Attack (DOS):** The attacker makes some computing or memory resources very busy or too full by doing some calculations to handle the legitimate logical request, denies legitimate users from accessing the machine.
2. **Probing Attack:** The attacker attempts to collect information about the computer network for a specific purpose by circumventing security controls.
3. **Remote to Local Attack (R2L):** This type of attack occurs when an attacker exploits vulnerabilities to provide local access to a network, and the attacker begins to send packets to the device over the network.
4. **User to Root Attack (U2R):** An attacker exploits root access, and an attacker could exploit some vulnerabilities to access a system's regular user account. Table 2 shows the attack types in the KDD99 dataset with the main attack category.

Table 2. KDD99 attack types.

Main attack	Attack type
Normal	normal
DOS	smurf, teardrop, neptune, back, pod, land
Probe	ipsweep, portsweep, nmap, satan
R2L	phf, guess_passwd, spy, warezmaster, ftp_write, warezclient, imap, multihop
U2R	buffe_overflow, loadmodule, perl, rootkit

Most researchers used KDD99 dataset to evaluate results and because of the computational requirements for the full KDD99 dataset and the inherent drawbacks of the dataset, mostly the researchers relied on part of the dataset and were trained and tested the model proposed. Here are some studies that used KDD99 dataset:

Othman et al.[21] proposed Spark-Chi-SVM model for intrusion detection using KDD99 dataset. ChiSqSelector is applied for feature Selection and the SVMwithSGD classifier is applied to build an intrusion detection model using Apache Spark. The results showed that the proposed model achieved high performance compared with the Chi-Logistic Regression classifier. The Spark-Chi-SVM experimental model showed high performance and less training time.

Lv et al. [22] proposed the KPCA-DEGSA-HKELM approach using a 10% subset of the KDD99 dataset and the UNSW-NB 15 dataset, which has been divided into the training and testing set. To reduce the dimensions and feature extraction, the Kernel Principal Component Analysis (KPCA) was applied. A combination of the differential evolution (DE) and gravitational search algorithm (GSA) is used to optimize the parameters of HKELM (Extreme Learning Machine with a Hybrid Kernel Function), which develops its global and local optimization abilities during prediction attacks. Then, KPCA-DEGSA-HKELM approach is obtained with achieved high accuracy and the time-saving.

Farooq et al. [23] used the NS-3 simulator and SVM classifier to determine whether the network traffic is normal or specific attack (Dos, Probe, R2L, and U2R) using KDD99 Dataset for training and testing. In the experiment, the authors used feature selection techniques. Results obtained showed the accuracy of 99.

Singh et al. [24] proposed a hybrid intrusion system (H-IDS) using the KDD99 dataset. H-IDS introduced a hybrid strategy with intelligent water drops to execute the feature selection (IWD) and support vector machine (SVM) for classification network traffic. Experimentations showed H-IDS helps to achieve the goal by attaining high classification, detection, and precision.

Ghasemi et al. [25] suggested a GA-KELM approach for built models on KDD99 and NSL-KDD datasets based on five different labels, have been gathered as a new dataset. GA used for feature selection task. Kernel Extreme Learning Machine (KELM) used as a classification algorithms. The proposed approach can easily outperform general classification algorithms which use all the features of the employed dataset with the highest accuracy.

3.2 NSL-KDD dataset

The NSL-KDD is a public dataset, which has been developed from the previous KDD99 dataset [19]. A statistical analysis performed on KDD99 dataset raised important issues that significantly affect the accuracy of intrusion detection and lead to a misleading evaluation of AIDS [26]. The main problem with KDD99 dataset is the huge amount of duplicate packets, analysis of training and testing KDD99 dataset revealed that approximately 78% and 75% of network packets are repeated in both training and test set [27]. Table 3 shows statistics of redundant instances in KDD99 train and test set.

Table 3. Statistics of redundant instances in KDD99 train set.

	Original instances	Distinct instances	Reduction Rate
Attacks	3,925,650	262,178	93.32%

Normal	972,781	812,814	16.44%
Total	4,898,431	1,074,992	78.05%

Statistics of redundant instances in KDD99 test set.

	Original instances	Distinct instances	Reduction Rate
Attacks	250,436	29,378	88.26%
Normal	60,591	47,911	20.92%
Total	311,027	77,289	75.15%

This huge amount of duplicate instances will affect the training set on machine learning methods to be biased to normal instances and thus prevent them from earning irregular instances which are usually more harmful to the computer system [27, 28].Although this new version of KDD99 dataset but it still has some problems and may not typically represent current real networks, due to lack of public datasets for network-based IDS, it can still be applied as an effective benchmark dataset to help researchers to evaluate different intrusion detection methods [29]. The advantages for NSL-KDD dataset are:

1. No redundant instances in the train set, so the classifier will not produce any biased result.
2. No duplicate instances in the test set which have better reduction rates.

NSL-KDD testing dataset consists of 22,544 instances and the training dataset consists of 125,973 instances.The size of NSL-KDD dataset is sufficient to make it practical to use the whole NSL-KDD dataset without the need for random sampling. NSL-KDD training and testing set instances are shown in table 4 with its class [30].

Table 4. NSL-KDD dataset instances.

Training dataset		Testing dataset	
Class	**Instances**	**Class**	**Instances**
Normal	67343	Normal	9711
DOS	45927	DOS	7458
Probe	11656	Probe	2421
R2L	995	R2L	2754
U2R	52	U2R	200
Total	125973	Total	22544

The 42 features include data about the various five classes of network connection, and each instance classifies as a normal class or into one of four attacks. The four classes are grouped as Dos, Probe, R2L, and U2R. The training dataset consists of 23 classes and the testing dataset consists of 38 classes that include 21 attacks from training dataset, 16 novel attacks and 1 normal class, class label of instances in the dataset are categorized into 5 main categories (Normal, Dos, Probe, U2R, and R2L). This dataset includes a large number of features to classify different attack types. The nature of features in NSL-KDD dataset is divided into four groups (Basic, Traffic, Host, and Content features). The types information of all the 41 features available in NSL-KDD dataset: four are binary, three are nominal, and 34 features are continuous [31]. Table 5 displays types of features in NSL-KDD dataset.

Table 5. Types of features in NSL-KDD dataset.

Type	Features
Nominal	protocol_type, Service, Flag.
Binary	land, logged_in, is_host_login, is_guest_login.
Numeric	duration, src_bytes, dst_bytes, wrong_fragment, urgent, hot,

num_failed_logins, num_compromised, root_shell, su_attempted, num_root, num_file_creations,num_shells, num_access_files, num_outbound_cmds, count, srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate
diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate.

Most researchers used NSL-KDD dataset to evaluate, mostly the researchers relied on part of the dataset and were trained and tested the model proposed. Here are some studies that used NSL-KDD dataset:

Bhati et al. [32] anal analyzed Linear SVM, Quadratic SVM, Fine Gaussian SVM, and Medium Gaussian SVM techniques on NSL-KDD dataset, which separated into two sets: one is a training set and another is testing. This analysis concluded that Fine GaussianSVM provides the best accuracy and least error for intrusion detection.

Biswas et al. [33] proposed an IDS model using five-fold cross-validation on NSL-KDD dataset. The authors used a different mix of feature selection algorithms and classifiers. IGR, PCA, CFS, and minimum redundancy maximum- relevance feature selection techniques are applied for feature selection. K-NN, DT, NN, SVM and NB classifiers are used for classifiers. K-NN classifier produced better performance than others and, among the feature selection methods, the IGR feature selection method is better than others.

Belavagi et al. [34] discussed the prediction analysis of different supervised machine learning algorithms namely Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and Random Forest using NSL-KDD dataset. Experimental results showed that the Random Forestachieved very good performance in identifying Dos, Probe, and U2R attacks, but it was poor in the identification of R2L attacks.

Thaseen et al. [35] suggested model for the intrusion detection using NSL-KDD dataset. Chi-square is applied for feature selection and multi class SVM is used as a classifier. Experimental results showed that the proposed model better in detection rate and reduced false alarm rate.

3.3 Kyoto 2006+ dataset

This dataset has been built on 3 years through honeypots data of Kyoto University [36, 37]. Therefore there is no manual labeling and anonymity process, but it has a bounded view of network traffic because only directed attacks on honeypots can be observed [38]. This dataset covers over three years of real traffic data collected from honeypots which were captured from Nov. 2006 to Aug. 2009 and regular servers that are deployed at Kyoto University [39]. During the observation period, there were 43,043,255 attack sessions, 425,719 unknown attacks sessions, and 50,033,015 normal sessions. Table 6 displays the overall characteristics of honeypot data in the Kyoto 2006+ dataset.

Table 6. Overall characteristics of honeypot data.

	Number of sessions	Average number of sessions per day
Total	93,076,270	93,638
Normal	50,033,015	50,335

Known attack	42,617,536	42,874
Unknown attack	425,719	428

Since normal traffic is frequently simulated during attacks and only produce DNS and mail traffic data, which does not reflect normal traffic in the real world, there are no false positives, which are important to reduce the number of alerts. This dataset consists of 24 statistical features: 14 conventional features and 10 additional features. Among them, the first 14 features were extracted based on KDD99 dataset, which is a very popular and widely used performance evaluation data for intrusion detection research. In addition to these 14 features, they have extracted 10 additional features that may enable them to more effectively investigate what is happening on their network. Of course, it can also be used for training and testing with 14 convention features [40]. This dataset is also available for Big Data analysis, of which size is 19.683 gigabytes. This dataset contains three class types: -1 attack, -2 shellcode, and 1 normal [41]. Kyoto 2006+ dataset features are shown in table 7.

Table 7. Features of Kyoto 2006+ dataset.

Feature Type	Feature
Conventional features	Duration, Service, Source_bytes, Destination_bytes, Count, Same_srv_rate, Serror_rate, Srv_serror_rate, Dst_host_count, Dst_host_srv_count, Dst_host_same_src_port_rate, Dst_host_serror_rate, Dst_host_srv_serror_rate, Flag (14).
Additional features	IDS_detection, Malware_detection, Ashula_detection, Label, Source_IP_Address , Source_Port_Number, Destination_IP_Address, Destination_Port_Number, Start_Time, Duration (10)

There are researchers used Kyoto 2006+ dataset to evaluate. Here are some studies that used Kyoto 2006+ dataset:

Kumar et al. [42] proposed Network Anomaly Detection Algorithm (NADA) based on distance measure and Relief-F. The proposed algorithm used KDD99 and Kyoto 2006+ datasets on Matlab. Common classification algorithms such as Naïve Bayes, SVM, and Decision Trees were also implemented. NADA outperforms all the other classifiers with regard to the time taken for execution. Experimental results observed that the detection rate, accuracy, F-Score, and MCC are higher in NADA and false alarm rate is lower.

Salo et al. [43] suggested the IG-PCA-Ensemble approach on three datasets, namely NSL-KDD, Kyoto 2006+, and ISCX 2012. The proposed model with the ensemble exhibited achieved better performance regarding false alarm rate, detection rate, and classification accuracy.

Sahu et al. [44] used the Decision Tree (J48) algorithm to classify the network packet. They used a labeled network dataset called Kyoto 2006+ dataset. For training and testing, they used 134665 network instances. Experimental he experimental results showed, the proposed model is able to detect unknown attacks.

3.4 ISCX 2012 dataset

Information Security Centre of Excellence (ISCX) was generated ISCX 2012 dataset by the Canadian Institute for cybersecurity [45]. ISCX 2012 was generated by a dynamic approach and present good guidelines for generating realistic and useful IDS evaluation

datasets during one week [46]. Their approach consists of: 1) the Alpha Profile has implemented various scenarios of multistage attacks to flow the abnormal segment of the dataset. 2) the beta profile is the benign traffic generator, produced realistic network traffic with background noise [47].

ISCX 2012 benchmark dataset contains statistical features (time_stamp, source_bytes, dst_bytes, source_packets, dst_packets, protocol, direction, Tag, source_ip, dst_ip) taken with a single interface on the switch to which all traffic is directed to it. In this dataset, the effect of real network traffic traces were analyzed to determine the normal behavior of computers from the real traffic of HTTP, IMAP, SMTP, POP3, SSH, and FTP protocols. It depends on realistic network traffic, which is labeled and contains various attack scenarios. It is a labeled dataset, comprises over two million traffic packets that attack data representing 2% of the whole traffic [48]. This dataset has four types of attack scenario consisting of Infiltrating the network from inside, Brute force SSH, HTTP denial of service (DoS), and Distributed Denial of service using an IRC botnet (DDoS) [49, 50]. Different attack scenarios are executed at different times and each attack consists of 5 steps: (1) information gathering and reconnaissance (passive or active), (2) vulnerability identification and scanning, (3) gaining access and compromising a system, (4) maintaining access and creating backdoors (5) and covering tracks.

The total size of ISCX 2012 dataset is 90.9 GigaBytes (GB). The traces were obtained in seven days of recent and realistic malicious and normal network activities under practical and systematic conditions [51]. Table 8 summarizes the complete ISCX 2012 dataset. As can be seen in Table8, every attack scenario was applied for only a single day and two days contained only regular traffic and explain the diversity of the regular network behavior and the complexity of the attack scenarios [48, 52].

Table 8. Overview of ISCX2012 dataset.

Date	Number of Flows	Number of Attacks	Description
11/6/2010 Friday	474,278	0	Normal Activity No malicious activity
12/6/2010 Saturday	133,193	2,086	Normal Activity Non-classified attacks
13/6/2010 Sunday	275,528	20,358	Infiltering the network from Inside Normal Activity.
14/6/2010 Monday	171,380	3,776	HTTP Denial of Service Normal Activity.
15/6/2010 Tuesday	571,698	37,460	Distributed Denial of Service using an IRC Botnet.
16/6/2010 Wednesday	522,263	11	Normal Activity No malicious activity.
17/6/2010 Thursday	397,595	5,219	Brute Force SSH Normal Activity.
Total	2,545,935	68,910	2.71% malicious

Note: “Number of attacks” is the subset of flows that contain an attack.

Although ISCX 2012 dataset includes real-life network attacks, it also has some shortcomings: A considerable amount of network flows was unlabeled, attack scenarios are not described in detail in terms of when the attack is started and ended and some flow records are given in uniflowformat whereas others are in biflow format and some flow records include null values [50]. When compared with the recent datasets this dataset can be characterized as follows: realistic network configuration because of the real testbed, realistic traffic because of the real and recent attacks [45].This dataset is provided in PCAP as well as a custom XML file for network flows created with the IBM QRadar device. The

XML flow file contains round truth labels, remember that network flow is collected from a number of IP packets and consists of source and destination IP addresses, source, destination port numbers, and protocol [52]. Table 9 summarizes the 14 features that can be extracted from the labeled XML file of network flows.

Table 9. ISCX 2012 dataset features.

No.	Feature	Description	Unique
1	SrcIP	Source IP address	2,478
2	DstIP	Dest. IP address	34,552
3	SrcPort	Source port	64,482
4	DstPort	Dest. port	24,238
5	AppName	Application name	107
6	Direction	Direction of flow	4
7	Protocol	IP protocol	6
8	Duration	Flow duration	N/A
9	TotalSrcBytes	Total source bytes	N/A
10	TotalDstBytes	Total dest. bytes	N/A
11	TotalBytes	Total bytes	N/A
12	TotalSrcPkts	Total source packets	N/A
13	TotalDstPkts	Total dest. packets	N/A
14	TotalPkts	Total packets	N/A

Note: “uniques” means the number of possible values of a categorical feature.

The some studies that used ISCX 2012 + dataset summarized as:

Mighan et al. [53] suggested a hybrid scheme that combines the advantages of a deep network and machine learning algorithms on Apache Spark. The autoencoder network used for feature extraction, which is followed by several classification such as support vector machine, random forest, decision trees, and Naive Bayes. The ISCX 2012 dataset is used in an experiment to validate the proposed model and evaluated the performance in terms of accuracy, f-measure, sensitivity, precision, and time.

Dwivedi et al. [54] proposed the EFSAGOA approach by using ISCX 2012 dataset. The EFS is used to rank the features for selecting the high ranked subset of features, and the AGOA is used to determine significant features. AGOA used the Support Vector Machine (SVM) as a fitness function to choose the extremely efficient features and to maximize the classification performance. The proposed approach obtained a high detection rate, accuracy, and low false alarm rate.

Aldwairi et al. [55] Restricted Boltzmann Machine technique (RBM) was applied to distinguish between normal and anomalous NetFlow traffic. RBM can be classified as normal and anomalous NetFlow traffic using ISCX 2012 dataset.

3.5 UNSW-NB 15 dataset

Comment [I9]: Format, Add space before

The UNSW-NB 15 dataset was generated in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) by the IXIA Storm tool to extract a hybrid of modern normal and modern attack behaviors[56]. It is one of the recent datasets to evaluate NIDS, it has become available to researchers since late 2015 [57].
A tcpdump tool was used to capture 100 GigaBytes (GB) of the raw network traffic (pcap files), each pcap file contains 1000 MB in order to make analysis of packets easier [58].

The simulation period was 16 hours on Jan 22, 2015, and 15 hours on Feb 17, 2015, for capturing 100 GB [59]. Twelve algorithms and tools such as Argus and Bro-IDS were executed in a parallel implementation to UNSW-NB15 dataset. It consists of 49 features and 2, 540,044 instances which are stored in four CSV files [60].The features of the UNSW-NB 15 dataset are categorized into six broad groups, the descriptions of which are given in Table 10.

Table 10. UNSW-NB 15 dataset features categorization.

No	Name of the category	Description
1	Flow features	It contains identifier attributes between hosts such as client-to-serve or server to-client.
2	Basic features	It includes features that distinguish the protocol connections.
3	Content features	It contains the TCP / IP features and also contains some features of the http services.
4	Time features	It contains of time features such as round trip time of TCP protocol start/end packet time arrival time between packets etc.
5	Additional generated features	Special purpose features that take care of service protocols.
	General purpose features(from number 36 - 40)	
	Connection features (from number 41- 47)	Built based on a chronological order of the last time feature.
6	Labelled Features	It represents the label of the instances.

The features are categorized into six groups that include (13) basic features, (8) content features, (9) time features, (7) connection features, (12) additional features and two features for class label. A total of 49 features determining the features of connections are present for each data instance. The features are mixed in nature with some being nominal, some being numeric (Integer, Binary and Float) and some taking on timestamp values as given in Table 11 [11].

Table 11. Features type of UNSW-NB15 dataset.

No.	Feature Type	Count
1	Nominal	6
2	Integer	28
3	Binary	3
4	Float	10
5	Timestamp	2

The dataset has a total number of 2540044 labeled instances, each being labeled either normal or attack, the total number of attacks in the dataset is 321283 instances and the total number of normal instances is 2218761. The size of the normal information packets represents 88% of the dataset size, while the attack information packets represent 12%. The distribution of instances across the two groups is presented in Table 12.

Table 12. Details of instances in UNSW-NB15 dataset.

Name	Count
Total Number of events	2540044

Normal	2218761
Attacks	321283

UNSWNB15 is a complex dataset, it represents modern network and attack traffic and can be used for reliable evaluation of NIDS[58]. The main categories of instances are nine types of attacks and one group representing the normal instances in the dataset. The attacks are categorized as Fuzzers, Reconnaissance, Shellcode, Analysis, Backdoors, DoS, Exploits, Generic, and Worms [59, 60]. The attacks, subcategory of attacks, and the distribution of all UNSW-NB 15 dataset instances are given in Table 13.

Table 13. Categorizations of attacks in UNSW-NB 15 dataset.

Attack type	Attack Subcategory	Number of Events
Normal	-	2,218,761
Fuzzers	FTP,HTTP,RIP,SMB,Syslog,PPTP,FTP,DCERPC, OSPF,TFTP,DCERPC,OSPF,BGP	24246
Reconnaissance	Telnet, SNMP, SunRPCPortmapper (TCP) UDP Service , SunRPCPortmapper (TCP) UDP Service, SunRPCPortmapper (TCP) TCP Service, SunRPCPortmapper (UDP) UDP Service, NetBIOS, DNS, HTTP, SunRPCPortmapper (UDP), ICMP, SCTP, MSSQL,SMTP,NETBIOS, DNS	13987
Shellcode	FreeBSD, HP-UX, NetBSD, AIX, SCO Unix, Linux, Decoders, IRIX, OpenBSD, Mac OS X, BSD, Windows, BSDi, Multiple OS, Solaris	1511
Analysis	HTML,Portscanner,Spam	2677
Backdoors	-	2329
DoS	Ethernet, Microsoft O_ce, VNC, IRC, RDP, TCP, VNC, FTP, LDAP, Oracle, TCP, TFTP, DCERPC, XINETD, IRC, SNMP, ISAKMP, NTP, Telnet, CUPS, Hypervisor, ICMP, SunRPC, IMAP, Asterisk, Browser	16353
Exploits	Evasions, SCCP, SSL, VNC, Backup Appliance, Browser, Clientside Microsoft O_ce, Interbase, Miscellaneous Batch, SOCKS, TCP, Apache,IMAP, Microsoft IIS, Clientside, Clientside Microsoft Paint, IDS, SSH, ICMP, IDS, DCERPC, FTP, RADIUS, SSL, WINS, POP3, Unix r Service, Cisco IOS, lientside Microsoft Media Player, Dameware,LPD,MSSQL ,O_ce Document, RTSP,SCADA,VNC, ebserver, All,LDAP, NNTP, IGMP, Oracle, RDesktop, Telnet, Apache, PHP, SMB, SunRPC, Web Application, DNS, Evasions, ADIUS, BrowserFTP, PPTP, SCCP,SIP,TFTP	44525
Generic	All,SIP, HTTP, SMTP, IXIA, TFTP, SuperFlow, HTTP, TFTP	215481
Worms	-	174

There are several recent studies that used UNSW-NB 15 dataset such as:

Thaseen et al. [61] proposed a correlation-based feature selection integrated with neural network for identifying anomalies attacks using NSL-KDD and UNSW-NB 15 dataset. The

results showed that the proposed model is superior in terms of accuracy, sensitivity, and specificity in comparison with other studies.

Nawir et al. [62] suggested a distributed online implementation of averaged one dependence estimator (DOAODE) method for a NIDS. They extended the prior work to predict the multi-class labels based on the UNSW-NB15 dataset [63]. The experimental results showed that the DOAODE classifier for is high in accuracy and fast to train the network traffic.

Raman et al.[64] Designed an intelligent IDS consists of an efficient feature selection technique and a robust classification model. The experimental validation used NSL-KDD and UNSW-NB 15 datasets under two scenarios: SVM trained with all features and SVM trained with optimal model features obtained from HC-IBGSA proved the significance of HC-IBGSA in terms of various performance metrics (classification accuracy, detection rate, and false alarm rate). The proposed HC-IBGSA SVM was implemented using python. The Weka and Matlab were used for validation purposes. The experimental displayed HC-IBGSA improved the performance of SVM in terms of false alarm rate and detection rate.

Belouch et al [65] evaluated the performance using four classification algorithms (SVM, Naïve Bayes, Decision Tree, and Random Forest) on Big Data processing tool. The general performance comparison evaluated in terms of training time, prediction time, and detection accuracy. The Random Forest classifier gave the best performance in terms of accuracy, sensitivity, specificity, and execution time.

3.6 CIDDS-001 dataset

The CIDDS-001 (Coburg Intrusion Detection DataSet) is a labeled flow-based dataset. This dataset developed for the evaluation purpose of Anomaly-based Network Intrusion Detection System (NIDS) [66]. CIDDS-001 dataset consists of unidirectional NetFlow data, it consists of traffic data from OpenStack environment having internal servers (backup, mail, file, and web) and External Servers External Server (file synchronization and web server), which is deployed on the internet to capture real-time and up-to-date traffic from the internet [67]. CIDDS-001 dataset consists of realistic normal and attacks traffic that allow for an important measurement of NIDS in a Cloud environment. It is divided into four parts each is created during a week. It contains 14 features, the first 10 features are the default NetFlow features and the last four features are additional features [68]. The CIDDS-001 dataset contains about 16 million flows. It was captured over a period of two weeks [69]. Attack flows are captured in the dataset within four attacks types (suspicious, unknown, attacker, and victim) [70, 71] . Table 14 provides a description for CIDDS-001 dataset features.

Table 14. CIDDS-001 dataset features.

No	Feature Name	Feature Description
1	Src IP	IP Address of the source node.
2	Src_Port	Port of the source node.
3	Dest_IP	IP Address of the destination node.
4	Dest_Port	Port of the destination node.
5	Proto	Transport Protocol (e.g. ICMP, TCP, or UDP).

6	Date_first_seen	Start time flow first seen.
7	Duration	Flow duration.
8	Bytes	Number of transmitted bytes.
9	Packets	Number of transmitted packets.
10	Flags	OR concatenation of all TCP Flags.
11	AttackDescription	Provides additional information about the set attack parameters (e.g. the number of attempted password guesses for SSH-Brute-Force attacks).
12	AttackType	Types of attack (portScan, dos, bruteForce, PingScan).
13	AttackID	Unique Attack id. Allows attacks which belong to the same class carry the same attack id.
14	Class	Class label (Normal, Attacker, Victim, Suspicious, and Unknown).

A lot of studies are being done on the development of effective NIDS using CIDDS-001 dataset. Here are some studies that used CIDDS-001 dataset:

Rashid et al. [72] introduced a comparative analysis on benchmark datasets NSL-KDD and CIDDS-001 using machine and deep learning algorithms. For getting optimal results, they used the hybrid featureselection and ranking methods. Six classification algorithms usedsuch as SVM, Naïve Bayes, k-NN, Neural Networks, DNN, and DAE. The experimental results showed that k-NN, SVM, NN, and DNN classifiers achieved high performance on the NSLKDD dataset whereas k-NN and Naïve Bayes classifiers achieved high performance on the CIDDS-001 dataset.

He et al. [73] suggested ensemble approach for feature selection on KDD99, UNSW-NB15, and CIDDS-001 datasets. They used Mean Decrease Impurity (MDI), Random Forest Classifier (RFC), Stability Selection (SS), Recursive Feature Elimination (RFE), and Chi-square to get the score of each feature. Then, a simple voting method used to integrate feature selection methods. Decision Tree (DT), k-NN, SVM, and Multi-Layer Perception (MLP) are used for classification. They compared the feature subsets with classification accuracy before and after the ensemble. The experiment showed that the EFS achieved high accuracy in classification.

Verma et al. [74] discussed the statistical analysis and evaluation using the CIDDS-001 dataset. Two techniques, k-nearest neighbor and k-means clustering were used. On the basis of evaluation results, it concluded that both k-nearest neighbor and k-means clustering perform well over CIDDS-001 dataset.

3.7 CICIDS2017 dataset

The CICIDS2017 provided by Canadian Institute for cybersecurity Intrusion Detection System, it is a very recent dataset [75]. CICIDS2017 contains up-to-date network attacks but also it meets all criteria of real-world attacks, it is a refinement of ISCX2012 dataset [45]. Since the start of CICIDS2017 dataset, the dataset has begun to attract researchers to analyze and develop new models and algorithms [76].
This dataset consists of labeled network flows, and including CSV files for machine and deep learning (MachineLearningCSV.zip) are publicly available for researchers and the corresponding profiles, full packet payloads in PCAP format, and the labeled flows (GeneratedLabelledFlows.zip) [77]. The machine learning file of the CICIDS2017 dataset (MachineLearningCSV.zip contains eight CSV files that represent the profile of the network traffic for five days, which includes normal and attack traffic for each day. This dataset contains attack information as five days traffic data, Thursday and Friday working hour

afternoon data are well suited for binary classification, Likewise, Tuesday, Wednesday, and Thursday morning data for designing a multiclass detection model [3]. The files containing of CICIDS-2017 dataset are displayed in table 15.

Table 15. Descriptions of files containing CICIDS-2017 dataset.

Name of Files	Attacks found	Flow count
Monday-WorkingHours.pcap_ISCX.csv	No Attack	529918
Tuesday-WorkingHours.pcap_ISCX.csv	Benign, FTP-Patator, SSH-Patator	445909
Wednesday-workingHours.pcap_ISCX.csv	Benign, DoSGoldenEye, DoS Hulk, DoSSlowhttptest, DoSslowloris, Heartbleed	692703
Thursday-WorkingHours-Morning-WebAttacks.pcap_ ISCX.csv	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS	170366
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ ISCX.csv	Benign, Infiltration	288602
Friday-WorkingHours-Morning.pcap_ISCX.csv	Benign, Bot	191033
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Benign, PortScan	225745
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Benign, DDoS	286467

However, it should be noted that the best detection model should be able to detect attacks of any type. Therefore, to design such as typical IDS, the traffic data of all day should be combined to form a single dataset to be used by IDS. The dataset shape in terms of the number of instances 2830743 and 79 features. The overall characteristics of CICIDS2017 dataset are shown in table 16 [78].

Table 16. Overall characteristics of CICIDS2017 dataset.

Dataset Name	CICIDS2018
Dataset Type	Multi class
Year of release	2017
Total number of distinct instances	2830743
Number of features	79
Number of distinct classes	15

According to the author of CICIDS2017 dataset, it stored in eight different files containing five days normal and attacks traffic data of the Canadian Institute of Cybersecurity [76, 78]. The dataset shape in terms of the number of 79.

CICIDS2017 dataset contains a wide range of attack types based on the 2016 McAfee report (DOS, DDOS, Web-based, Brute force, Infiltration, Scan, Bot, and Heart-bleed), it is publicly available. The whole shape of a dataset that contains 2830743 instances and 79 features (78 features plus one for attacks type labels) containing 15 class labels (1 normal and 14 attacks). Surprisingly, no redundant instances found. The characteristics of the CICIDS2017dataset and the detailed class occurrence are displayed in table 17.

Table 17. Class instance occurrence of CICIDS2017 dataset.

Class Labels	Flow count
BENIGN	2273097
DoS Hulk	231073
PortScan	158930
DDoS	128027
DoSGoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoSslowloris	5796
DoSSlowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
Web Attack – XSS	652
Infiltration	36
Web Attack – Sql Injection	21
Heartbleed	11
Total	2830743

The some studies that used CICIDS2017 dataset are:

Krishna et al. [79] introduced Fast k-Nearest Neighbor Classifier (FkNN) as a better machine learning algorithm for NIDS on Cloud Environment. From the experimental results, they concluded that the FkNN classifier achieved high accuracy with less detection time.

Alrowaily et al. [80] applied seven machine learning algorithms using CICIDS2017 dataset. They used several performance metrics to examine the algorithms. The experimental results displayed that the K-Nearest Neighbors (KNN) classifier outperformed in terms of accuracy, recall, precision, and F1-score as compared to other classifiers.

Zhang1 et al. [81] proposed a real-time detection system for high-speed network environments, which is implemented by a distributed Random Forest classification algorithm based on Apache Spark. They implemented using CICIDS2017 dataset. The experimental results and comparisons showed that the proposed detection model has a shorter detection time, achieved higher accuracy, and can realize a real-time intrusion detection in a high-speed network environment.

3.8 CSE-CIC-IDS2018 dataset

The CSE-CIC-IDS2018 created by Communications Security Establishment (CSE) and Canadian Institute for Cybersecurity (CIC) in 2018 for intrusion detection and malware anticipation, datasets by CIC and ISCX have been utilized worldwide [82]. Furthermore, the dataset was enhanced by considering the criteria used to create the CIC-IDS201 [83]. It contains different attack scenarios: DoS, DDoS, Heartbleed, Brute-force, Botnet, Web attacks, and inside network infiltration. The attacking infrastructure includes 50 machines and the victim organization has 5 divisions and includes 420 machines and 30 servers [84].

This dataset has been published online for researchers with nearly 5 million data in CSV and PCAP format. The unprocessed PCAP data should be used if new features need to be extracted. The CSV format dataset can be used in artificial intelligence technologies.
The dataset was edited daily, and raw data were recorded. When creating data, 80 statistical properties such as time, number of packets, number of bytes, packet length, etc. The numbers of attacks and number of instances are shown in table 18[85].

Table 18. Class instance occurrence of CSE-CIC-IDS2018 dataset.

Class	Number of Instances
Benign	2,856,035
Bot	286,191
Brute Force	513
DoS	1,289,544
Infiltration	93,063
SQL injection	53
Total	4,525,399

There are several recent studies that used CSE-CIC-IDS 2018 dataset such as:

Karatas et al. [83] applied six machine learning IDS (Decision Tree, K Nearest Neighbor, Gradient Boosting, Random Forest, Adaboost, and Linear Discriminant Analysis algorithms) by using CSE-CIC-IDS2018 dataset, it is an imbalanced dataset. To reduce the imbalance, Synthetic Minority Oversampling TEchnique (SMOTE) was applied. The use of the dataset from which samples were taken increased the average resolution of the samples. The experimental results demonstrated that the implemented models have very good accuracy.
Kanimozhi et al. [86] proposed Artificial Neural Networks by using an up-to-date cybersecurity dataset (CSE-CIC-IDS2018). The proposed approach provided an outstanding performance of Accuracy and average area under the ROC (Receiver Operator Characteristic) curve, and the average False Positive rate.

Kim et al. [85] suggested Convolutional Neural Network (CNN) model and Recurrent Neural Network (RNN) model using KDD99 and CSE-CIC-IDS2018 datasets. The experimental results displayed the CNN model was able to identify DoS attacks compared to the RNN model.
Lin et al. [84] proposed a dynamic network anomaly detection system using CSE-CIC-IDS2018 dataset. They used LSTM to build the neural network model and incorporate the attention mechanisms to deal with time-correlated network traffic classification issues. In order to solve the class-imbalance problem, they used the SMOTE algorithm as well as the improved loss function to optimize the training process. The experimental results achieved a very good result in traffic classification.

Table 19 summarizes the above benchmark datasets. In table 19, we order the datasets from the recent to oldest. It displays the number of instances and features, names of attacks in each dataset, and if the available data set is divided into two files, one for training and the other for testing.

Table 19. Public datasets for NIDS.

Dataset	Number of features	Number of instances	Name of attacks	Separate train-test set
CSE-CIC-IDS2018 [82]	80	4,525,399	Bot, Brute Force,	No

Comment [l10]: **Should add more details for the table like comparisons**

-Separate train-test set (which Methods can be using K-fold or SMOTH,)

-Which can be using unlabeled like an unsupervised or labeled supervisor
- which one potential to used for Anomaly, Signature, Hybrid
- which when can be got more Performance, Accuracy and False positive Rate FPR, and can say more reliability.

			Dos, Infiltration, SQL injection.	
CICIDS2017 [77]	79	2830743	DoS Hulk, PortScan, DDoS, DoSGoldenEye, FTP-Patator, SSH-Patator, DoSslowloris, DoSSlowhttptest, Bot, Web Attack – Brute Force, Web Attack – XSS, Infiltration, Web Attack – Sql Injection, Heartbleed.	No
CIDDS-001 [69]	14	16 million	suspicious, unknown, attacker, and victim	No
UNSW-NB15 [87]	49	2540044	Fuzzers, Reconnaissance, Shellcode, Analysis, Backdoors, DoS, Exploits, Generic, and Worms	Yes
ISCX2012 [88]	14	2,545,935	Infiltrating, Brute force SSH, HTTP denial of service (DoS), and Distributed Denial of service (DDoS).	No
Kyoto 2006+ [89]	24	93,076,270	Attack, Shellcode.	No
NSL-KDD [26]	42	148,517	Dos, Probe, R2L, and U2R.	Yes
KDD99 [90]	42	4,898,431	Dos, Probe, R2L, and U2R.	Yes

4. DISCUSSION AND RECOMMENDATION

Comment [l11]: Add more details to clear for recommendation and highlight of which dataset better and explain why and if there some barriers or gab, highlighted that for readers of the paper that main objective of your work

Network-based datasets are essential for NIDS training and evaluation. It can be used to compare the quality of different NIDS with each other. In any case, the datasets must be represented to be suitable for those tasks. The community is aware of the importance of realistic network data. Therefore, this paper analyzed public available datasets in NIDS to support researchers to find the appropriate dataset for their specific evaluation scenario. Furthermore, this work focuses on a collection of dataset properties as a basis for comparing available datasets and for identifying suitable datasets.

There are different approaches that have been used for improving NIDS efficiency using machine learning algorithms and publically available datasets. A detailed analysis of publically available datasets is introduced to help researchers to keep research time and find an appropriate dataset of NIDS. The recommendations in this paper have been coming from our analysis of eight datasets. The authors make the following recommendations about the use of available datasets:

- **Use recent datasets:** As mentioned above, no perfect dataset exists for NIDS. However, this paper demonstrates that there are many datasets available for packet and flow-based network traffic. So, we recommend users to evaluate their intrusion detection methods with more than one dataset to avoid overfitting to a particular dataset and evaluate their methods in a more general context. Moreover, this paper recommends users to use recent datasets such as UNSW NB15, CIDDS-001, CICIDS2017, and CSE-CIC-IDS2018 in evaluating NIDS; it reflects modern scenarios of attacks.
- **Use of several datasets:** Therefore, there could be another approach for using more than one dataset and real-world network traffic to emphasize these points. To ensure reproducibility for third parties, the authors recommend evaluating intrusion detection approaches using at least one publicly available dataset.
- The general recommendation to use CICIDS2017, CIDDS-001, CSE-CIC-IDS2018, and UNSW-NB 15 datasets. These datasets may be suitable for general evaluation settings. CICIDS2017, UNSW-NB 15, and CSE-CIC-IDS2018 datasets contain a wide range of attack scenarios.
- The recommendation does not sight that other datasets are inappropriate. For instance, KDD99, Kyoto 2006+, NSL-KDD, and ISCX2012 datasets do not include in our recommendation due to their increasing age.
- Only publicly available can be used by third parties. Thus it do as the base for NIDS evaluation. Furthermore, the quality of the datasets can only be checked by third parties if they are publicly available.
- The authors recommend the publication of additional metadata for the CSE-CIC-IDS2018 dataset such that third parties can analyze the data and their results in more detail.

Comment [I12]: Why recommended, Explain by example more clear, if like add some features, or what??

5. CONCLUSION

Network-based datasets are essential for training and evaluating intrusion detection methods. This paper introduced a detailed analysis of benchmark and recent datasets for network intrusion detection systems. The authors described eight well-known datasets that include: KDD99, NSL-KDD, KYOTO 2006+, ISCX2012, UNSW-NB 15, CIDDS-001, CICIDS2017, and CSE-CIC-IDS2018. For each dataset, we provided a detailed analysis of it is instances, features, classes, and the nature of the features. The main objective of this paper was to offer an overview of the available datasets for NIDS and what each dataset was consists of. Furthermore, it presented some recommendations for using benchmark network-based datasets. As future work, it is possible to work in enhancing the current work by implementing various machine learning algorithms using recent datasets.

Comment [I13]: Write what your found, should mention

Comment [I14]: Clear more this considering for contribution, that statement still need clear

REFERENCES

1. Rathore, M.M., A. Ahmad, and A. Paul, *Real time intrusion detection system for ultra-high-speed big data environments.* The Journal of Supercomputing, 2016. **72**(9): p. 3489-3510.
2. Tavallaee, M., et al. *A detailed analysis of the KDD CUP 99 data set*. in *2009 IEEE symposium on computational intelligence for security and defense applications*. 2009. IEEE.

3. Panigrahi, R. and S. Borah, *A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems.* International Journal of Engineering & Technology, 2018. **7**(3.24): p. 479-482.
4. Khraisat, A., et al., *Survey of intrusion detection systems: techniques, datasets and challenges.* Cybersecurity, 2019. **2**(1): p. 20.
5. ABDULRAHEEM, M.H. and N.B. IBRAHEEM, *A DETAILED ANALYSIS OF NEW INTRUSION DETECTION DATASET.* Journal of Theoretical and Applied Information Technology, 2019. **97**(17).
6. Abdulrazaq, M. and A. Salih, *Combination of multi classification algorithms for intrusion detection system.* Int. J. Sci. Eng. Res., 2015. **6**(1): p. 1364-1371.
7. Chitrakar, R. and C. Huang. *Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and naive bayes classification*. in *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing*. 2012. IEEE.
8. Othman, S.M., et al., *Survey on Intrusion Detection System Types.* International Journal of Cyber-Security and Digital Forensics, 2018. **7**(4): p. 444-463.
9. Ring, M., et al., *A survey of network-based intrusion detection data sets.* Computers & Security, 2019. **86**: p. 147-167.
10. Rani, N. and R.K. Purwar, *Performance Analysis of various classifiers using Benchmark Datasets in Weka tools.* International Journal of Engineering Trends and Technology (IJETT), 2017. **47**(5).
11. Hamid, Y., et al., *Benchmark Datasets for Network Intrusion Detection: A Review.* IJ Network Security, 2018. **20**(4): p. 645-654.
12. Alshamy, R. and M. Ghurab, *A Review of Big Data in Network Intrusion Detection System: Challenges, Approaches, Datasets, and Tools.* Journal of Computer Sciences and Engineering, 2020. **8**(7): p. 62-74.
13. Ferrag, M.A., et al., *Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study.* Journal of Information Security and Applications, 2020. **50**: p. 102419.
14. Hindy, H., et al., *A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems.* IEEE Access, 2020.
15. Chapaneri, R. and S. Shah, *A comprehensive survey of machine learning-based network intrusion detection*, in *Smart Intelligent Computing and Applications*. 2019, Springer. p. 345-356.
16. Viegas, E.K., A.O. Santin, and L.S. Oliveira, *Toward a reliable anomaly-based intrusion detection in real-world environments.* Computer Networks, 2017. **127**: p. 200-216.
17. Kaja, N., A. Shaout, and D. Ma, *An intelligent intrusion detection system.* Applied Intelligence, 2019. **49**(9): p. 3235-3247.
18. Özgür, A. and H. Erdem, *A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015.* PeerJ Preprints, 2016. **4**: p. e1954v1.
19. Chandolikar, N. and V. Nandavadekar. *Efficient algorithm for intrusion attack classification by analyzing KDD Cup 99*. in *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*. 2012. IEEE.
20. Kushwaha, P., H. Buckchash, and B. Raman. *Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99*. in *TENCON 2017-2017 IEEE Region 10 Conference*. 2017. IEEE.
21. Othman, S.M., et al., *Intrusion detection model using machine learning algorithm on Big Data environment.* Journal of Big Data, 2018. **5**(1): p. 34.
22. Lv, L., et al., *A novel intrusion detection system based on an optimal hybrid kernel extreme learning machine.* Knowledge-Based Systems, 2020: p. 105648.
23. Farooq, M.U., H. Xiaoli, and S.A. Rauf, *Big Data Security Analysis in Network Intrusion Detection System.* International Journal of Computer Applications, 2020. **975**: p. 8887.

24. Singh, P. and M. Venkatesan. *Hybrid approach for intrusion detection system*. in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. 2018. IEEE.
25. Ghasemi, J., J. Esmaily, and R. Moradinezhad, *Intrusion detection system using an optimized kernel extreme learning machine and efficient features.* Sādhanā, 2020. **45**(1): p. 1-9.
26. *NSL-KDD*. 2009; Available from: https://www.unb.ca/cic/datasets/nsl.html.
27. McHugh, J., *Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory.* ACM Transactions on Information and System Security (TISSEC), 2000. **3**(4): p. 262-294.
28. Gao, X., et al., *An adaptive ensemble machine learning model for intrusion detection.* IEEE Access, 2019. **7**: p. 82512-82521.
29. Revathi, S. and A. Malathi, *A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection.* International Journal of Engineering Research & Technology (IJERT), 2013. **2**(12): p. 1848-1853.
30. Verma, P., et al. *Network intrusion detection using clustering and gradient boosting*. in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2018. IEEE.
31. Dhanabal, L. and S. Shantharajah, *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): p. 446-452.
32. Bhati, B.S. and C. Rai, *Analysis of Support Vector Machine-based Intrusion Detection Techniques.* Arabian Journal for Science and Engineering, 2019: p. 1-13.
33. Biswas, S.K., *Intrusion detection using machine learning: A comparison study.* International Journal of Pure and Applied Mathematics, 2018. **118**(19): p. 101-114.
34. Belavagi, M.C. and B. Muniyal. *Multi Class Machine Learning Algorithms for Intrusion Detection-A Performance Study*. in *International Symposium on Security in Computing and Communication*. 2017. Springer.
35. Thaseen, I.S. and C.A. Kumar, *Intrusion detection model using fusion of chi-square feature selection and multi class SVM.* Journal of King Saud University-Computer and Information Sciences, 2017. **29**(4): p. 462-472.
36. Song, J., et al. *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*. 2011.
37. Sato, M., H. Yamaki, and H. Takakura. *Unknown attacks detection using feature extraction from anomaly-based ids alerts*. in *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*. 2012. IEEE.
38. Gharib, A., et al. *An evaluation framework for intrusion detection dataset*. in *2016 International Conference on Information Science and Security (ICISS)*. 2016. IEEE.
39. Song, J., H. Takakura, and Y. Okabe. *Cooperation of intelligent honeypots to detect unknown malicious codes*. in *2008 WOMBAT Workshop on Information Security Threats Data Collection and Sharing*. 2008. IEEE.
40. Song, J., H. Takakura, and Y. Okabe, *Description of kyoto university benchmark data.* Available at link: http://www. takakura. com/Kyoto_data/BenchmarkData-Description-v5. pdf [Accessed on 15 March 2016], 2006.
41. Park, K., Y. Song, and Y.-G. Cheong. *Classification of attack types for intrusion detection systems using a machine learning algorithm*. in *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2018. IEEE.
42. Kumar, D.A. and S. Venugopalan, *A design of a parallel network anomaly detection algorithm based on classification.* International Journal of Information Technology, 2019: p. 1-14.

43. Salo, F., A.B. Nassif, and A. Essex, *Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection.* Computer Networks, 2019. **148**: p. 164-175.
44. Sahu, S. and B.M. Mehtre. *Network intrusion detection system using J48 Decision Tree*. in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2015. IEEE.
45. Shiravi, A., et al., *Toward developing a systematic approach to generate benchmark datasets for intrusion detection.* computers & security, 2012. **31**(3): p. 357-374.
46. Kumar, G., *An improved ensemble approach for effective intrusion detection.* The Journal of Supercomputing, 2020. **76**(1): p. 275-291.
47. Sharafaldin, I., A.H. Lashkari, and A.A. Ghorbani. *Toward generating a new intrusion detection dataset and intrusion traffic characterization*. in *ICISSP*. 2018.
48. Mighan, S.N. and M. Kahani. *Deep learning based latent feature extraction for intrusion detection*. in *Electrical Engineering (ICEE), Iranian Conference on*. 2018. IEEE.
49. Sallay, H., et al. *A real time adaptive intrusion detection alert classifier for high speed networks*. in *2013 IEEE 12th International Symposium on Network Computing and Applications*. 2013. IEEE.
50. Pektaş, A. and T. Acarman, *A deep learning method to detect network intrusion through flow-based features.* International Journal of Network Management, 2019. **29**(3): p. e2050.
51. Khan, M.A., M. Karim, and Y. Kim, *A scalable and hybrid intrusion detection system based on the convolutional-LSTM network.* Symmetry, 2019. **11**(4): p. 583.
52. Fernández, G.C. and S. Xu. *A Case Study on Using Deep Learning for Network Intrusion Detection*. in *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. 2019. IEEE.
53. Mighan, S.N. and M. Kahani, *A novel scalable intrusion detection system based on deep learning.* International Journal of Information Security, 2020: p. 1-17.
54. Dwivedi, S., et al., *Implementation of adaptive scheme in evolutionary technique for anomaly-based intrusion detection.* Evolutionary Intelligence, 2020. **13**(1): p. 103-117.
55. Aldwairi, T., D. Perera, and M.A. Novotny, *An evaluation of the performance of Restricted Boltzmann Machines as a model for anomaly network intrusion detection.* Computer Networks, 2018. **144**: p. 111-119.
56. Moustafa, N. and J. Slay, *A hybrid feature selection for network intrusion detection systems: Central points.* arXiv preprint arXiv:1707.05505, 2017.
57. Faker, O. and E. Dogdu. *Intrusion detection using big data and deep learning techniques*. in *Proceedings of the 2019 ACM Southeast Conference*. 2019.
58. Moustafa, N. and J. Slay, *The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set.* Information Security Journal: A Global Perspective, 2016. **25**(1-3): p. 18-31.
59. Moustafa, N. and J. Slay. *UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)*. in *2015 military communications and information systems conference (MilCIS)*. 2015. IEEE.
60. Moustafa, N. and J. Slay. *The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems*. in *2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS)*. 2015. IEEE.
61. Sumaiya Thaseen, I., et al., *An integrated intrusion detection system using correlation-based attribute selection and artificial neural network.* Transactions on Emerging Telecommunications Technologies, 2020: p. e4014.
62. Nawir, M., et al. *Distributed Online Averaged One Dependence Estimator (DOAODE) Algorithm for Multi-class Classification of Network Anomaly Detection System*. in *IOP Conference Series: Materials Science and Engineering*. 2019. IOP Publishing.

63. Nawir, M., et al. *Performances of machine learning algorithms for binary classification of network anomaly detection system*. in *Journal of Physics: Conference Series*. 2018.
64. Raman, M.G., et al., *An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm.* Artificial Intelligence Review, 2019: p. 1-32.
65. Belouch, M., S. El Hadaj, and M. Idhammad, *Performance evaluation of intrusion detection based on machine learning using Apache Spark.* Procedia Computer Science, 2018. **127**: p. 1-6.
66. Ring, M., et al. *Flow-based benchmark data sets for intrusion detection*. in *Proceedings of the 16th European conference on cyber warfare and security*. 2017.
67. Verma, A. and V. Ranga, *On evaluation of network intrusion detection systems: Statistical analysis of CIDDS-001 dataset using machine learning techniques.* Pertanika Journal of Science & Technology, 2018. **26**(3).
68. Idhammad, M., K. Afdel, and M. Belouch, *Distributed intrusion detection system for cloud environments based on data mining techniques.* Procedia Computer Science, 2018. **127**: p. 35-41.
69. *CIDDS-001*. 2017; Available from: https://www.hs-coburg.de/forschung-kooperation/forschungsprojekte-oeffentlich/ingenieurwissenschaften/cidds-coburg-intrusion-detection-data-sets.html.
70. Althubiti, S.A., E.M. Jones, and K. Roy. *Lstm for anomaly-based network intrusion detection*. in *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*. 2018. IEEE.
71. Tama, B.A. and K.-H. Rhee, *Attack classification analysis of IoT network via deep learning approach.* Res. Briefs Inf. Commun. Technol. Evol.(ReBICTE), 2017. **3**: p. 1-9.
72. Rashid, A., M.J. Siddique, and S.M. Ahmed. *Machine and Deep Learning Based Comparative Analysis Using Hybrid Approaches for Intrusion Detection System*. in *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*. 2020. IEEE.
73. He, W., H. Li, and J. Li. *Ensemble Feature Selection for Improving Intrusion Detection Classification Accuracy*. in *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*. 2019.
74. Verma, A. and V. Ranga, *Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning.* Procedia Computer Science, 2018. **125**: p. 709-716.
75. Singh Panwar, S., Y. Raiwani, and L.S. Panwar. *Evaluation of Network Intrusion Detection with Features Selection and Machine Learning Algorithms on CICIDS-2017 Dataset*. in *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*. 2019.
76. Nicholas, L., et al., *Study of long short-term memory in flow-based network intrusion detection system.* Journal of Intelligent & Fuzzy Systems, 2018. **35**(6): p. 5947-5957.
77. *CICIDS2017* 2017; Available from: http://www.unb.ca/cic/datasets/IDS2017.html.
78. Panwar, S.S., et al., *Implementation of Machine Learning Algorithms on CICIDS-2017 Dataset for Intrusion Detection Using WEKA.*
79. Krishna, K.V., K. Swathi, and B.B. Rao, *A Novel Framework for NIDS through Fast kNN Classifier on CICIDS2017 Dataset.*
80. Alrowaily, M., F. Alenezi, and Z. Lu. *Effectiveness of machine learning based intrusion detection systems*. in *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*. 2019. Springer.
81. Zhang, H., et al. *Real-time distributed-random-forest-based network intrusion detection system using Apache spark*. in *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*. 2018. IEEE.

82. *CSE-CIC-IDS2018* 2018; Available from: https://www.unb.ca/cic/datasets/ids-2018.html.
83. Karatas, G., O. Demir, and O.K. Sahingoz, *Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset.* IEEE Access, 2020. **8**: p. 32150-32162.
84. Lin, P., K. Ye, and C.-Z. Xu. *Dynamic network anomaly detection system by using deep learning techniques*. in *International Conference on Cloud Computing*. 2019. Springer.
85. Kim, J., et al., *CNN-Based Network Intrusion Detection against Denial-of-Service Attacks.* Electronics, 2020. **9**(6): p. 916.
86. Kanimozhi, V. and T.P. Jacob. *Artificial intelligence based network intrusion detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing*. in *2019 International Conference on Communication and Signal Processing (ICCSP)*. 2019. IEEE.
87. *UNSW-NB 15*. 2015; Available from: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-
Datasets/.
88. *ISCX 2012*. 2012; Available from: https://www.unb.ca/cic/datasets/ids.html.
89. *Kyoto2006+*. 2006; Available from: http://www.takakura.com/Kyoto_data/.
90. *KDD99*. 1999; Available from: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.