A Survey of Advancement in AnomalyIntrusion Detection System

Archana Gondalia (archana.gondalia@gujgov.edu.in)
L.D. College of Engineering
Dr. Apurva Shah
Maharaja Sayajirao University of Baroda

Research Article

Keywords: Anomaly Detection, Intelligent neural Cybercrime and Cyber Attack Intrusion Detection

Posted Date: November 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2284207/v1

Additional Declarations: No competing interests reported.

A Survey of Advancement in Anomaly Intrusion Detection System

Mrs. Archana Gondalia[1*] and Dr. Apurva Shah[2†]

[1*]Department of Computer Engineering, L. D. College of Engineering, 120, Circular Road, University Area, Ahmedabad, 380015, Gujarat, India.

[2]Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Kalabhavan, Rajmahal Road, Nr. Kirtistambh, Vadodara, 390001, Gujarat, India.

*Corresponding author(s). E-mail(s): archana.gondalia@gmail.com;
Contributing authors: apurva.shah-cse@msubaroda.ac.in;
†These authors contributed equally to this work.

Abstract

Every day, trillions of data transfer takes place on the internet. With such huge data transfer the hackers are evolving new and anomalous techniques to intrude and misuse it. Different neural approaches were implemented for the Intrusion Detection System (IDS) based on deep learning (DL) and machine learning (ML) frameworks that helped to maximize the forecasting accuracy. Researchers to help identify intrusion detection upto significant accuracy. However, because of the processing of a huge volume of data having redundant characteristics with irrelevant features, the efficiency of the IDS model is reduced. The researchers use a variety of feature selection strategies to avoid processing of irrelevant and redundant features. Selection of proper features leads to improvement in detection rate as well as processing time. This survey paper aims to provide insight on utilization of various data sets namely KDD Cup'99, NSL-KDD, Kyoto 2006+, UNSW-NB15, Canadian Institute for Cybersecurity Intrusion Detection System (CICIDS) 2017, Aegean WiFi Intrusion (AWID), Australian Defense Force Academy (ADFA), Cambridge and University of Brescia (UNIBS), Communications-Security Establishment and the Canadian-Institute for Cybersecurity (CSE-CIC) IDS 2018 in IDS. This survey

paper also describes various classifiers and matrices used for anomaly intrusion detection. The key objective of the present research work to improve dataset for the identification of accurate intrusion detection.

Keywords: Anomaly Detection, Intelligent neural Cybercrime and Cyber Attack Intrusion Detection

1 Introduction

One of the most important considerations is security, issue for all the networks in today's environment. Intruders and hackers have done many attempts successfully to down company network and web services. Network security includes of many fields of research like cyber security, cyber forensics, risk management, system log analysis, Intrusion Detection System (IDS), securing access control mechanism etc. The data generated from network needs to ensure confidentiality, integrity and availability. For this network security plays the role. There have been numerous ways established to keep the network safe connection and configuration over the internet. Various network tools are used to study network activities like Nmap, Nessus, Wireshark, Snort etc. [1] To execute network attacks the intruder excerpts the required information from the apprehended network packets. The network attacks either conducted within network, organization or outside network, organization and these attacks are known as internal network attacks and external network attacks respectively. Moreover, the IDS has employed to forecast these attacks. Also, IDS can help to identify malicious activity of network user without conceding the security of network and the host. Intrusion detection system studies may could be measured as a classification task which distinguishes among malicious and normal characteristics in the network. IDS may could be categorized in a variety of ways in the basis of gathered database, the data studied, and the actions that must be taken. In addition, It can be categorized in dual ways based on where it is installed in the network based IDS (NIDS) and Host-based IDS (HIDS). For devices on the enterprise network, a HIDS were monitored and has analyzed the system's installed applications and its configurations. On the other hand, the HIDS sensors have been placed on any gadgets including servers or any computer desktop [2] With the use of NIDS sensors, monitoring and network collision analysing for suspicious functions and other malicious events. Before transmitting the data, all packets are stored in the cloud storage header frame. Basically IDS can be classified like anomaly and signature based or detection of intrusions based on misuse in terms of recognition or detection methods. Pre-defined patterns of intrusion are kept on database so that specific type of attack can be represented by each pattern in misuse or signature based detection method. Thus, to look for patterns that are similar to patterns that is present in dataset which can identify known intrusion in signature based intrusion detection method [3]. If no patterns exist in the database, the

network will be unable to detect new attacks. Like result, IDS has been validated by estimating the false alarm score and exactness score for malicious prediction. Moreover, the rules for the malicious forecasting process are made in the basis of the network prediction behavior. So the normal network behavior model once builds then each packet which violates the condition will be intrusion. In addition, the rule based IDS are suitable for forecasting the different and unknown attacks. As this type of detection is difficult to distinct among malicious and authenticated behavior of network that was gained the highest false alarm measure. Table 1 represents IDS classification.

IDS Taxonomy	Source of Information	Methods for detection
HIDS	✓	
NIDS	✓	
Source of Information	✓	
Anomaly Detection		✓
Signature Detection		✓

Table 1: IDS classification

Various methods implemented to detect intrusions such as classification using semi-supervised learning, supervised learning, un-supervised learning, neural network, rule mining, and other DL based approaches. To ensure security of network these techniques were implemented by identifying attacks with a high degree of precision. An appropriate techniques for application-specific datasets, must be determined to calculate performance of an IDS. With the study of different features selected of the chosen dataset, the classification and detection can be performed. The hybrid feature selection strategy was described in [4]. Here, ChiSqSelector was processed the feature selection function and the intrusion specification has been defined based on regression and support vector models to build platform capable of high-speed data analysis. The suggested method analyses data by selecting suitable attributes that can help categorise attacks. A method that is based on C4.5, RF, Naïve-Bayes (NB), and REPTree algorithm implemented in [5] with selection for features. Here, Information Gain (IG) and Correlation based methods implemented for features selection by assigning weight or rank to the feature. Many researchers used bio-inspired algorithms with DL or Machine Learning (ML) method to select feature or optimization with parameters.

2 Related Work

IDS is a computer or network security system that detects unusual activity and missing signatures. To enhance the effectiveness of IDS, several methodologies and frameworks have been developed, including semi-supervised, unsupervised, and supervised machine learning methods. In [6], convolutional networks , autoencoders, together with recurrent networks were employed as training data

with different NSLKDD datasets. A comparison of anomaly detection algorithms based on deep learning and well-known classification systems such as RF, SVM, Decision-Tree, K-nearest model, and extreme learning is also provided. Using test data that has never been seen before and common reliability of classification indicators such as the mean Average Precision (mAP), RoC Curve, Area under the RoC, and Accuracy-Recall Curve these models were compared to each other and to conventional machine learning models and accuracy of classification.

The IEEE 802.11 standard has been the target of the most prevalent attacks in AWID family of data set evaluated in [7] with Random Forest and J48 classification algorithms for detecting intrusions in wireless networks. In [8], proposed a model named Spark-Chi-SVM to detect IDS. On Apache Spark is a big data platform developed by Apache, the author employed ChiSqSelector for feature selection and SVM classifier to introduced the different features based IDS. Different Regression classifiers were compared. The outcome of the experiment revealed that the implemented approach takes less time to set up train data, also, it's good for Big Data RNN used to detect intrusion with a deep learning technique was proposed in [9]. Together with this author, we investigated the model's performance in dual and multi features neurons, which maximized the learning rate of the categorization, as well as the effects of the number of neurons and learning rate on the proposed model's performance. Here, the wide anomaly detection approach has been executed, which is processed in the basis of multi-step outliers and technique to select feature that can efficiently find a associated optimal clust6ering subset features, like tree based datasets [10]. This IDS is more suitable for the network application. Hence, to find the shortest route optimal path search algorithm has been implemented in the clustering model [11]. Also, the historical malicious features dictionary is employed to detect the threat activities in the network, for the verification digital signature [12]. The siginificant components of IDS is the ability to identify the source of the incursion is feature selection. [13] on the basis of each score feature decided upon during the selection procedure, sought by eliminating non-relevant features and discover characteristics that will have an impact for improving the rate of detection. [14] to locate the exemplars from the audit data, researchers utilised dual level clustering models that based in K-means and Affinity Propagation [12] through historical data analysis, aimed to construct a network profile called DSNSF that signifies the projected typical behaviour of a network traffic activity.

[15] Proposed a method for selection of feature using mutual information can handle non-linearly and linearly dependent data features and has been tested in network intrusion detection scenarios. As per the proposed feature selection approach one new method created for IDS. In comparison to the original 6-dimensional dataset, the experimental findings of CNN outperform the k-NN and SVM classifiers, delivering a reduced false alarm rate together with higher detection rates and accuracy. [16] massive data processing for intrusion

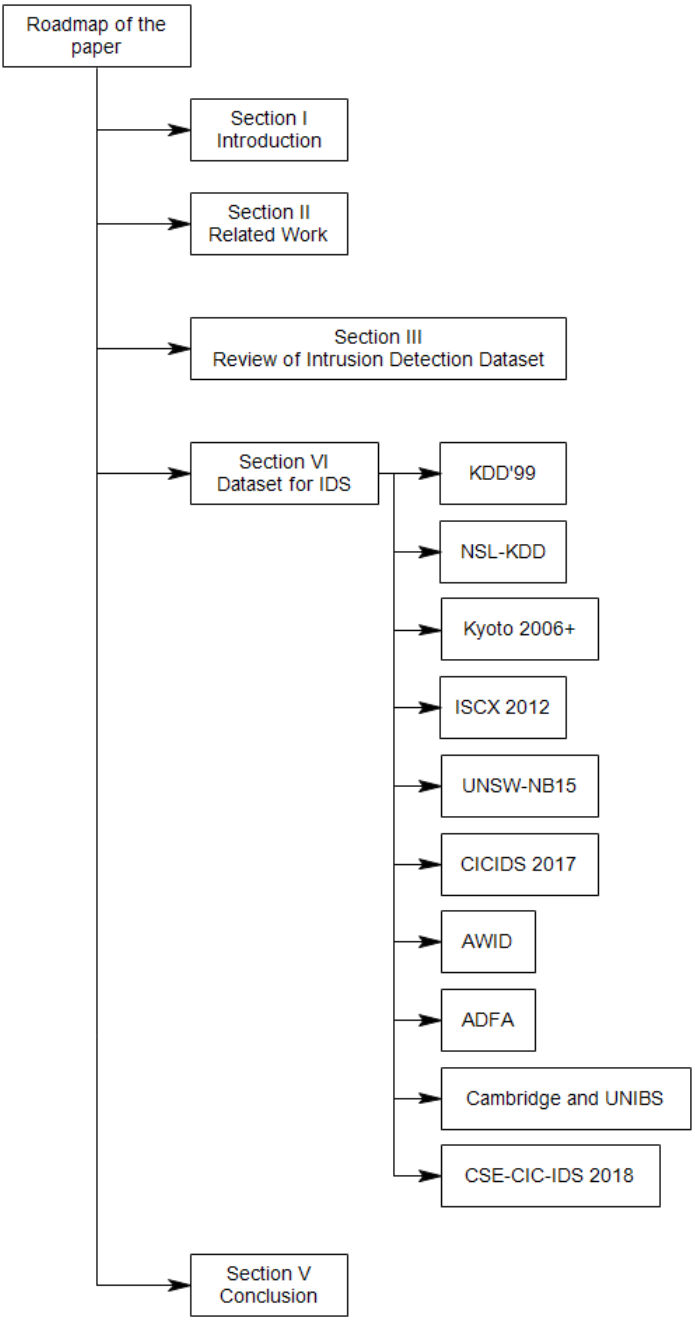

Fig. 1: Roadmap of the Paper

detection is a major challenge. In order to identify subsets of the characteristics based on network intrusion detection KDD'99 data, multiple strategies were also used [17]. As a dimensional reduction technique, a distributed deep belief network model is used, and based on Apache Spark an iterative reduce strategy for multi layer ensemble SVM is used to retrieve features which are informative in order to produce a more accurate prediction ensemble [18].

The cyber security domain's cornerstone is network intrusion detection. Despite the fact that dozens of studies have lately been conducted on network detected intrusion, rare to find a comprehensive study on the topic. In [4], a detailed empirical investigation utilising the following DL models for network intrusion detection is presented: Depp networks, gated Deep networks, deep memory model like long short term and Deep Belief Networks (DBN). Results with the investigations showed that each and every model of DL was effective in the detection of network threats, increasing accuracy and forecasting score by 5% to 10% and decreasing false-Alarm Rate (FAR) by 1% to 5% in most cases. The primary reason of a low FAR in detection is due to the presence of duplicated with unnecessary dataset attributes. [19]introduced a new technique to pick in one step, both attribute subset and hyperparameters are available, as well as three deep learning models to deal with this challenge. When compared to without the equivalent values of the same models pretraining on the similar data set, when deploying technique the experimental data demonstrate a considerable improvement in network intrusion detection with FAR decreasing by 1% to 5% and DR increasing by 4% to 6%.

In the realm of information assurance, preventing unknown malicious event and other harmful cyber threats are a difficult by conventional IDS [20]. Polymorphic mechanisms are used by intruders to disguise the attack payload and avoid detection. To enhance the effectiveness of IDS, several unsupervised and supervised procedures are implemented for the IDS based on the Ml principle. In [20], an unique semi-supervised learning strategy with fuzziness was employed for improving the detection features of the IDS by combining Using a supervised learning method, unlabeled samples were analyzed. To construct the addition layer in the deep feed neural network, fuzzy features have been trained. Moreover, the designed layers the hidden parameter for specifying the unlabeled samples. [21] proposes an an autonomous anomaly detection technique for large-scale, high-dimensional data and unlabeled data sets. The algorithm is a cross between a DBN and a one-class SVM. In [22], GA was utilised to develop an anomaly intrusion detection model and to classify the data SVM was utilised with known and unknown attacks for feature selection. In [3], the ABC mechanism is utilized to choose attributes, while ensemble features in the classification function has provided the rich classification outcome. In [5], four ML algorithms, namely C4.5, NB, RF, and REPTree, were used to identify the most essential features based on filter method on a re-sampled version of KDD'99. [23] proposes a new approach for binarizing a continuous pigeon inspired optimizer. With FPR, True Positive Rate (TPR), accuracy, F-score, the suggested approach beat numerous methods to choose feature from

relevant research. [24], ACO is used to find significant traits that are optimally selected. The road-map of the paper is in Figure 1 which represents the flow of the paper.

3 Review of Intrusion Detection Dataset

One dataset which reflect to real world networks with a repository of sufficient amount consistent data can be considered as standard database. To value the NIDS Performance, a large standard data is required. Researchers faced one of the challenges is to find a suitable dataset during assessment of intrusion detection. A common difficulty that cyber security research teams encounter is obtaining a dataset from the real world which depicts the network traffic traversing. As most of the institutions and companies who are utilizing the internet services will not allow data to be observed, shared or recorded is fact. Because of the laws and regulations are there associated to confidentiality and privacy this challenge faced by researcher. In the case when the data is

Sr No	Developed By	Dataset Name	Attack Types	Features
1	University of California	KDD CUP 99	Dos, R2L, U2R, Probe	41
2	University of California	NSL-KDD	Dos, R2L, U2R, Probe	41
3	Kyoto University	Kyoto	Normal and Attack sessions	24
4	University of New Brunswick	ISCX2012	DoS, DDoS, Bruteforce, Infiltration	IP flows
5	UNSW Canberra	UNSW-NB15	Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms	49
6	Canadian Institue of Cyber Security	CIC-IDS-2017	Brute force, Portscan, Botnet, Dos, DDoS, Web, Infiltration	80
7	University of New South Wales	AFDA	Zero-day attacks, Stealth attack, C100 Webshell attack	System call traces
8	Canadian Institue of Cyber Security	CSE-CICIDS-2018	Brute force, Portscan, Botnet, Dos, DDoS, Web, Infiltration	80

Table 2: Overview of IDS Datasets produced from actual network activity traces

accessible or shared to researcher then also lot many alteration with the data in which some of the portions altered, deleted or restricted. Due to this, a lot many important data which needs to be considered by the researchers will be lost and any inferences like statistical may no longer reliable. The table 2 shows the Overview of IDS Datasets produced from actual network activity traces.

4 DATA-SET FOR IDS

Since 1998, IDS databases have been created in enormous amount. In this paper, we have selected different IDS datasets, like, NSL-KDD, KDD-Cup'99, ISCX 2012, Kyoto 2006+, UNSW-NB15, CICIDS 2017, AWID, ADFA, Cambridge and UNIBS, CSE-CIC-IDS 2018 and each dataset explained briefly in this paper.

4.1 KDD'99

From the 1998 DARPA dataset with preprocessing the portion of tcpdump, dataset was termed as KDD-CUP'99 was made [25]. In recent, the data science field has utilized KDD'99 for the first time and it is now regarded a standard benchmark for assessing IDS [26]. The dataset is basically intended for the recognizing interruptions in an organization like military climate [23]. Following table 3 describes the number of elements present in Normal, DoS, Probe, R2L and U2R attacks:

Attacks	Training	Testing
DoS	391458	223298
R2L	1126	5993
Probe	4107	2377
U2R	52	39
Normal	97278	60593

Table 3: KDD'99 Dataset

In this dataset, there is combination of 41 additional features that contain the malicious features in each TCP connection [27]. There are 4 categories in dataset with 24 attacks and normal network traffic [28].

1. Denial of Service(DoS): To restrict network usage by interfering with the targeted users' ability to access services;
2. Probe: To get information, the attacker examines the system's network
3. Users to Root (U2R): Connect with a client xreditionals and attempts to get root access via system flaws;
4. Remote to Local (R2L): have a local account but are attempting to get root access

In testing dataset, total 38 attacks are there from that 14 attacks are not there in the training dataset. In addition, KDD-CUP'99 datasets have included three labelled sets, like training the whole data or ratio based training and testing. Total 494,021 instances are present in this database [29]. DoS attack consumes memory to check the behavior of legitimate user, Probe attack has scanned the info of the network and hack the communication process. Moreover, the R2L attacks have sent the additional packets for cause the data overhead issues. U2R attacks uploaded the threat behavior in the network

Ref	Applicability	Classification Algorithm	Dataset	Result
[5]	To improve accuracy and efficiency by selecting important features	C4.5,NB, RF and REPTree	KDD'99 Dataset	For DoS attack: DR: 99.6%, FPR: 0.3%, Probe Attack: DR: 99.8% FPR: 2.7%
[30]	A novel approch for intrusion detection	multi-layer SVM	KDD'99 Dataset	DR: 95.03%
[31]	To detect intrusion in the network	Neural Network: Self Organizing Maps	KDD'99 Dataset	Accuracy: 90%

Table 4: KDD'99 dataset for IDS

system [10]. Table 4 represents usage of KDD'99 dataset with classification algorithm and applicability for IDS.

4.2 NSL-KDD

The dataset NSL-KDD is used by researchers for IDS. This dataset, from the initial KDD CUP '99, is made up with selected records [32]. This set is an enhanced category of the KDDcup'99 dataset. There are some problems in the dataset that is related to attack data and synthesizing the network due to issues of privacy, unclear attack definitions and an unknown packet loss caused by network traffic [33]. So with the improvement from KDD'99 dataset, in this dataset the attacks are labelled and grouped according to how difficult they are to detect and duplicate records were removed. Hence a new dataset as a whole includes annotated entries from the KDD'99 dataset with no shortfalls encountered in NSL-KDD dataset. Around 10% of KDD'99 dataset included in this dataset as in KDD'99 dataset huge amount of information is present [34].

Ref	Applicability	Classification Algorithm	Dataset	Result
[35]	To generate a predictive model to determine exact attack categories	SVM, Naïve Bayes, Neural Network	NSL-KDD	Accuracy: 96%
[22]	To develop hierarchical intrusion detection model.	Fuzzy C-Means(FCM) with Genetic Algorithm and SVM	NSL-KDD	Accuracy: 99.76%, Detection Rate: 99.94%, False Alarm: 0.6%
[36]	developed a network intrusion detection system using an unsupervised learning algorithm	Neural Network: Self Organizing Maps	NSL-KDD	Accuracy: 90%

Table 5: NSL-KDD dataset for IDS

In this dataset, instances are considered either normal or attack. To define each connection, the dataset offers 41 attributes derived from network traffic [31]. From 41 features, 38 are numeric features and three non-numeric features (like flag, service and protocol type). Moreover, these features are specified as traffic, content based and basic features. Furthermore, the NSL-KDD has a label of class constricted to the subsequent mainly of four groups: DoS,U2R, R2L, Probe, Normal [37]. Table 5 represents usage of NSL-KDD dataset with classification algorithm and applicability for IDS.

4.3 Kyoto 2006+ Data set

At Kyoto University, real time traffic between Nov 2006 and Aug 2009 without any change from human Kyoto 2006+ dataset was built. Using email server, darknet sensors, honey pots, web crawler and email server dataset Kyoto 2006+ has been taken [38]. In the IDS, the malicious detection framework is based on traffic packet in formation. So, Kyoto 2006+ is characterized as consists of

Feature No.	Feature Name
KF1	Duration
KF2	service
KF3	Source bytes
KF4	Destination bytes
KF5	count
KF6	Same srv rate
KF7	Serror rate
KF8	Srvserror rate
KF9	Dst host count
KF10	Dst host srv count
KF11	Dst host same src port rate
KF12	Dst host serror rate
KF13	Dst host srvserror rate
KF14	Flag
KF15	Source Port Number
KF16	Destination Port Number
KF17	Label

Table 6: Kyoto University benchmark dataset

Ref	Applicability	Classification Algorithm	Dataset	Result
[15]	To achieve better accuracy and lower computational cost	Least Square Support Vector Machine based IDS (LSSVM-IDS)	Kyoto 2006+	DR: 99.64% FPR: 0.13% Accuracy: 99.77%
[40]	Dimensionality reduction for Network Intrusion Detection	SVM, IBK (k-NN), MLP	Kyoto 2006+	Accuracy: 98.95%, Detection Rate: 99.8%, FAR: 0.021%
[41]	Novel ensemble classifier (RFAODE) for intrusion detection system	RF Tree and Average One-Dependence Estimator (AODE)	Kyoto 2006+	Accuracy: 90.51% Detection Rate: 92.38% FAR: 0.14%

Table 7: Kyoto 2006+ dataset for IDS

43503 unique connections [39]. There are attributes 24 from which 14 attributes areinclined in the KDD-CUP 99 database with statistical information and other 10 features parameters were characterized in the form of typical flow-based features like ports, IP addresses, or duration of each packets transmission interval [39]. Other functions include IDS detection, label, malware detection, destination and source IP address, malicious detection, length, source port number, start time, and target port number, among others [40]. The presence of attack can be recognized by a label attribute. The Kyoto 2006+ data set contains around 93 million sessions over three years more than usual long period. All these features were explained in Table 6 in which features counts other than KF17, KF16, KF14, KF15 and KF2 are continuous and rest are categorical type [38]. Table 7 represents usage of Kyoto 2006+ dataset with classification algorithm and applicability for IDS.

4.4 The ISCX 2012 dataset

As one of the contemporary benchmark informational collection, the institute of information security has worked on the Intrusion Detection database[40]. In 2012 by catching traffic in an imitated network climate more than multi

week The ISCX data index was made [39]. The information contains in ISCX dataset around 85 GB of organization traffic information alongside profiles that portray the progression of the information and the assaults that happened during that week and was created utilizing genuine organization settings by catching parcels continuously all through a time of seven days [42].

Day	Friday	Monday	Sunday	Wednesday	Tuesday	Thursday	Saturday
Size	16.1 GB	6.85 GB	3.95 GB	17.6 GB	23.4 GB	12.3 GB	4.22 GB
Normal	3,78,667	1,67,609	2,75,170	5,22,263	5,34,320	3,92,392	1,33,111
Attack	0	3,771	20,358	0	37,378	5,203	2082
Ratio (Attack/Normal)	0	0.022499	0.073983	0	0.069954	0.01326	0.015641

Table 8: ISCX 2012 Data set

Data packets have been detected by human assistance with request to keep a strategic distance from any unwanted features of post-blending network attacks with constant foundation traffic for each day of the week [11]. For seven days of network activity, the dataset tracked live packets using protocols for different scenarios of malicious and normal activity in different network communication protocols that included transfer control and hyper link protocol [40]. The dataset includes of seven days catching with in general 2,450,324 traffic streams and it has 8,720 traffic streams [43]. This dataset contains 19 features [40]. Following reasons are there to download the ISCX dataset [44]. Total traffic flow records in this database is 2,450,324 and 19 attack behavior features. Different IP with seven days monitored real packets, and included both regular and malicious scenarios [40]. Data size, attack, normal flow and ratio of attack vs normal day wise of ISCX dataset shown in Table 8 [43]. Table 9 represents usage of ISCX dataset with classification algorithm and applicability for IDS.

Ref	Applicability	Classification Algorithm	Dataset	Result
[3]	Used to select the best subset of related features to detect network connections and because of the high ability of these algorithms	AdaBoost algorithms	ISCX 2012	Detection Rate:99.61% False Positive Rate: 0.01 Accuracy: 98.90
[45]	A hybrid model to detect anomaly and signature based intrusion detection	Spark ML(SVM, DT, RF, and Gradient Boosting tree (GBT) classifiers), Conv-LSTM	ISCX 2012	Accuracy: 97.29%
[40]	Dimensionality Reduction for Network Intrusion Detection	SVM, IBK (k-NN), MLP	ISCX 2012	Accuracy: 99.01% Detection Rate: 99.1% FAR: 0.01%

Table 9: ISCX 2012 dataset for IDS

4.5 UNSW-NB15 dataset

The database UNSW-NB15 was anticipated by [46], it has included the malicious events footprint scenarios and reflect the modern network traffic features. With the use of the IXIA tool at the Australian Cyber Security was generated synthetic anomalous network traffic and a selective descriptive of real modern normal behavior [47]. Tcpdump is a tool that can hold up to 100 GB of

Pcap data and can be used to simulate nine distinct sorts of attacks as well as generate real-world and contemporary threat models [23]. Records like binary, float, nominal, and integer features were built within 31 hours in UNSW-NB15 data with small emulated environment. There are 42 features, with normal class having 37,000 samples and anomaly class having 45,332 samples in the anomaly class in the full training dataset [48]. Total 9 malicious features are available in this specific database that are DoS, Generic, Exploits, Fuzzers, Reconnaissance, Analysis, Backdoors, Worms, and Shellcode are consists of real modern normal packets in the dataset [18]. [49]. Moreover, this dataset has been process in the ratio of 70% training and 30% testing[50]. A Model for Intrusion Detection System presented in [51] with genetic algorithm as feature extraction algorithm and Multiscale Convolutional Neural Network with Long Short-Term Memory (MSCNN-LSTM) as classification algorithm has been used. Total 45 different IP addresses are there in the dataset and is publicly available [52]. Table 10 represents usage of UNSW-NB15 dataset with classification algorithm and applicability for IDS. As compare to NSL-KDD dataset, the UNSW-NB15 data set has several advantages [18] which are as follows:

- The dataset comprises modern synthesized attack actions and real contemporary normal behaviors;
- The The training and testing sets are distributed similarly probability wise;
- To replicate the network packets competently, dataset includes a set of features from the payload and header of packets.;
- On existing classification systems, the complication of assessing the UNSW-NB15 is an indication of this data set's multifaceted patterns.

Ref	Applicability	Classification Algorithm	Dataset	Result
[48]	A hybrid feature selection and two-level classifier ensemble IDS	Rotation Forest, Bagging	UNSW-NB15	Accuracy: 91.27% FPR: 8.90% Sensitivity: 91.30%
[50]	Data optimization method to build IDS	Random Forest (RF)	UNSW-NB15	Accuracy: 92.8% FPR: 0.33%
[51]	Model for Intrusion Detection System	Multiscale Convolutional Neural Network with Long Short-Term Memory (MSCNN-LSTM)	UNSW-NB15	Accuracy: 95.6%

Table 10: UNSW-NB15 dataset for IDS

4.6 CICIDS2017

The CICIDS2017 is discovered in 2017 [1], it is standard data mostly utilized for the network application. This dataset is basically built for the researcher to evaluate models properly which is trustworthy and latest [19]. The dataset has captured total 5 days of data in which capturing period from 3rd of July, 2017 to 7th of July, 2017, started at 9:00 am and ended at 05:00 pm [18]. This CSV-formatted dataset demonstrates the usage of CICFlow Meters with labelled network traffic flow analysis depending on the ports of source and destination, IPs, time stamp, protocols, and attacks. Additionally, attacks like innocuous and the latest current popular assaults, which resemble actual data

Ref	Applicability	Classification Algorithm	Dataset	Result
[53]	To detect various types of attacks with high accuracy and efficiency.	C4.5, Random Forest (RF), and Forest by Penalizing Attributes (Forest PA) algorithms	CICIDS2017	Accuracy with binary classification: 99.89% Accuracy with multi-class classification:99.89%
[19]	Empirical study on network intrusion detection	Deep Neural Networks(DNN), Long Short-Term Memory Recurrent Neural Networks(LSTM-RNN), and Deep Belief Networks(DBN)	CICIDS2017	Accuracy with Binary Classification: DNN model: 97.85% LSTM-RNN: 98.83% DBN: 99.91% Accuracy with Multi-class Classification: DNN model: 88.04% LSTM-RNN: 92.41% DBN: 95.81%
[18]	For the detection of abnormal behavior in large-scale networks.	multi-layer ensemble SVM	CICIDS2017	Precision:90.40% Recall:95.65% F-measure: 92.95%

Table 11: CICIDS2017 dataset for IDS

packet sniffer. Moreover, the attacks such as Cross-area Scripting, SQL Injection, Incursion, distributed denial-of-service (DDoS), Port Scan, Brute Force, and Botnet, along with approximately 8 files containing 2,830,743 files and 78 distinct features with their label, are included in the most recent IDS datasets [53]. Attack-Network and Victim-Network are 2 different networks used in an environment infrastructure in the dataset.

1. The Victim-Network is utilised to give benign behaviour utilising the B-Profile system [14]
2. The Attack-Network is being subjugated to manage the effects of malicious events [19]

The two of them are set up with the basic PCs and organization gadgets running changed working frameworks. CICFlowMeter has been utilized to examine the apprehended pcap data over five working days for this dataset. Moreover, the digital communication and social sites are mostly depends of the networks access. Some social communication applications are, email, Facebook, twitter, etc. Total 20 attack flows which are grouped into categories like Botnet, DoS, DDoS, Heart bleed, Brute Force, Infiltration and Web attack CICIDS2017 carry unique 80 variables like class name to distinguish the specific overhead details and to find the causing occurrence. CICIDS2017 is an exceptionally tremendous dataset which has around 3 million organization streams in various documents as compare to NSL-KDD dataset, due to their predefined function and packages of testing and training cases. Table 11 represents usage of CICIDS2017 dataset with classification algorithm and applicability for IDS.

4.7 AWID Dataset

In 2015, AWID was released to the public as a collection of WiFi network sets information, with comprise authentic trace down both anomalous and normal acquired along with real-world organisation situations [54]. AWID is an openly accessible informational collection [39]. This dataset utilized in a little organization climate (11 customers) together with caught WLAN traffic in bundle based arrangement [54]. The AWID is variety of datasets which contains two equivalent set that is concede just on the marking technique (AWIDCLS,

AWID-ATK) [54]. Each quality has numeric or nominal values, and all data point in the dataset is provided as 155 vector ascribes [53]. Table 12 represents usage of AWID dataset with classification algorithm and applicability for IDS.The dataset can be categorised into two categories on the basis of number of target classes: AWID-CLS and AWIDATK.

- The AWID-CLS dataset is divided in four categories: impersonation, flooding, injection and normal [53]
- Four sets of primary class variables and three targets features are available in the AWID-ATK database [53]

Ref	Applicability	Classification Algorithm	Dataset	Result
[53]	To detect various types of attacks with high accuracy and efficiency.	C4.5, Random Forest (RF), and Forest by Penalizing Attributes (Forest PA) algorithms	AWID	Accuracy with binary classification: 99.52% Accuracy with multi-class classification:99.52%

Table 12: AWID dataset for IDS

4.8 ADFA Data set

For assessment of IDSs, the ADFA Linux (ADFA-LD) created by [55] cyber security benchmarks datasets. To collect this dataset Ubuntu Linux version 11.04 operating system has been used. To share files, a web server, to change network settings and movement of database syastems are different functions accessible by Ubuntu Linux configuration [33].

Ref	Applicability	Classification Algorithm	Dataset	Result
[33]	To identify both the well-known intrusions and zero-day attacks	C5, one class SVM	ADFA	Accuracy:97.40%

Table 13: ADFA dataset for IDS

For anomaly based systems the ADFA-LD12 is build and not for signature recognition IDS [56]. For offering web server Datasabse server, remote access and FTP server authors have installed Apache, MySQL and Tikiwiki to create ADFA dataset. Based on default ports FTP, SSH, secure web server and MySQL database are initiated [25]. This data set can be found on the internet for free and can be found in [15]. For HIDS evaluation, the dataset which is used to give dataset which is modern [33]. Table 13 represents usage of ADFA dataset with classification algorithm and applicability for IDS.

4.9 Cambridge and UNIBS

At the University of Cambridge, the computer laboratory published trace out traffic [57]. With comparison & assessment for the methods of classification for traffic, are the comprehensively suitable for traces [58]. Composition of this

traffic traces was done at the Genome Campus network in August, 2003. Ten different separation of these traffic traces collected in alternate time period of the 24-hour day [2]. To associate for web association, around 1000 users shrouded in this exploration and Gigabit Ethernet interface used for the same. The traffic was divided into ten blocks of around 28 minutes each to create ten datasets [59]. These data sets have 248 features for each flow sample, which are acquired using tcptrace from the transportation layer header [60]. Content-based or manual identification reliably identifies all flows, as stated in [57]. It is obvious that WWW traffic is much higher than other types of traffic [58]. Because it incorporates the FTP used for bulk data transfer, the BULK class always has the most bytes [59]. The telecommunication networks group of the University of Brescia [61] publishes UNIBS traffic traces.

Ref	Applicability	Classification Algorithm	Dataset	Result
[58]	A new feature optimization approach based on deep learning	C4.5, Decision Trees (C4.5), Support Vector Machine (SVM) and Naïve Bayes Kernel (NBK)	Cambridge and UNIBS	Flow OA: C4.5: 0.978±0.009 SVM: 0.984±0.00 NBK: 0.943±0.006 Byte OA: C4.5: 0.887±0.098 SVM: 0.920±0.039 NBK: 0.838±0.075 Flow g-mean: C4.5: 0.601±0.134 SVM: 0.511±0.139 NBK: 0.407±0.204 Byte g-mean: C4.5: 0.391±0.266 SVM: 0.120±0.097 NBK: 0.327±0.271
[59]	To identify a relevant feature subset for every class.	C4.5, Decision Tree	Cambridge and UNIBS	Flow Accuracy>96% Byte Accuracy>93%

Table 14: Cambridge and UNIBS dataset for IDS

Tcpdump [52] on the Faculty's router was used to capture two data sets in 2009 of September and October, respectively. The router is connected to Internet [58] by a dedicated 100 Mbps uplink. The router used a dedicated 100Mb/s uplink [59] to connect the campus network to the Internet. To characterise each flow, 35 statistical flow features were retrieved, including protocol, port number, time, bytes, and so on. Table 14 represents usage of Cambridge and UNIBS dataset with classification algorithm and applicability for IDS.

4.10 CSE-CIC-IDS 2018 Data set

The CIC-CSE was collaborated on analysing ids [62]. Brute-force, DDoS, Bot, Heartbled, DoS, Web threats, and penetration are among the seven attack scenarios included in the data set. The dataset contains system logs and network traffic, with that 83 attributes are extracted, including packet count, bytes count, duration and packet length [63].This data set's backward (destination-to-source) and forward (source-to-destination) paths are defined in the first packet. Table 15 represents usage of CSE-CIC-IDS 2018 dataset with classification algorithm and applicability for IDS.

Ref	Applicability	Classification Algorithm	Dataset	Result
[63]	An effective deep learning method for IDS	auto-encoder and K-means/GMM.	CSE-CIC-IDS 2018	Recall for Bot Dense Matrix: 96.98% and for Sparse Matrix: 99.73%

Table 15: CSE-CIC-IDS 2018 dataset for IDS

5 Conclusion

The study examines datasets created in the intrusion detection system (IDS) domain. These data sets utilised to assess the ML and DM based IDS's effectiveness. The study showed the necessity for an update in the basic dataset to more accurately determine current threats in the realm of IDS. This is due to the intruders carry out attacks utilising a variety of techniques and tools. Additionally, the methodology for carrying out various attacks replicate the requirement for datasets with accurate deployment scenarios. To meet the criterion of constructing an CIC-IDS-2017 and CSE-CICIDS are two intrusion detection datasets containing upgraded attacks and authentic network activity. Types of data for 2018 have already been released. This study examines the features, applicability, classification algorithm used with these datasets. In order to overcome the drawbacks of these datasets, we will concentrate on analysing how well they perform using different feature selection techniques, as well as adding ML and DM algorithms and data sampling.

6 Declarations

6.1 Ethical Approval

I will abide by the laws, rules, and regulations of the Journal. I will conduct myself with integrity, fidelity, and honesty. I will openly take responsibility for my actions, and only make agreements, which I intend to keep.

6.2 Availability of supporting data

The data that support the findings of this study are openly available in reference of the paper.

6.3 Competing interests

Not Applicable

6.4 Funding

Not Applicable

6.5 Authors' contributions

Both authors have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

6.6 Acknowledgements

We thank the anonymous referees for their useful suggestions.

References

[1] of New Brunswick, U.: Canadian Institute for Cybersecurity - IDS. 2017. https://www.unb.ca/cic/datasets/ids-2017.html. [Online; accessed 10-09-2021] (2021)

[2] WEKA: The workbench for machine learning. https://www.cs.waikato.ac.nz/ml/weka/ (2021)

[3] Mazini, M., Shirazi, B., Mahdavi, I.: Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and adaboost algorithms. Journal of King Saud University - Computer and Information Sciences **31**(4), 541–553 (2019). https://doi.org/10.1016/j.jksuci.2018.03.011

[4] Elmasry, W., Akbulut, A., Zaim1, A.H.: Empirical study on multiclass classification-based network intrusion detection. Computational Intelligence **35**, 919–954 (2019). https://doi.org/10.1111/coin.12220

[5] TCHAKOUCHT], T.A., EZZIYYANI, M.: Building a fast intrusion detection system for high-speed-networks: Probe and dos attacks detection. Procedia Computer Science **127**, 521–530 (2018). https://doi.org/10.1016/j.procs.2018.01.151. PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2017

[6] Naseer, S., Saleem, Y., Khalid, S., Bashir, M.K., Han, J., Iqbal, M.M., Han, K.: Enhanced network anomaly detection based on deep neural networks. IEEE Access **6**, 48231–48246 (2018). https://doi.org/10.1109/ACCESS.2018.2863036

[7] Kolias, C., Kambourakis, G., Stavrou, A., Gritzalis, S.: Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset. IEEE Communications Surveys Tutorials **18**(1), 184–208 (2016). https://doi.org/10.1109/COMST.2015.2402161

[8] Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T., Al-Hashida, A.Y.: Intrusion detection model using machine learning algorithm on big data

environment. Journal of Big Data (2018). https://doi.org/10.1186/s40537-018-0145-4

[9] Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access **5**, 21954–21961 (2017). https://doi.org/10.1109/ACCESS.2017.2762418

[10] Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: A multi-step outlier-based anomaly detection approach to network-wide traffic. Information Sciences **348**, 243–271 (2016). https://doi.org/10.1016/j.ins.2016.02.023

[11] Costa, K.A.P., Pereira, L.A.M., Nakamura, R.Y.M., Pereira, C.R., Papa, J.P., Falcão, A.X.: A nature-inspired approach to speed up optimum-path forest clustering and its application to intrusion detection in computer networks. Information Sciences **294**, 95–108 (2015). https://doi.org/10.1016/j.ins.2014.09.025. Innovative Applications of Artificial Neural Networks in Engineering

[12] Fernandes, G., Rodrigues, J.J.P.C., Proença, M.L.: Autonomous profile-based anomaly detection system using principal component analysis and flow analysis. Applied Soft Computing **34**, 513–525 (2015). https://doi.org/10.1016/j.asoc.2015.05.019

[13] Nkiama, H., Said, S.Z.M., Saidu, M.: A subset feature elimination mechanism for intrusion detection system. International Journal of Advanced Computer Science and Applications **7** (2016)

[14] Wang, W., Liu, J., Pitsilis, G., Zhang, X.: Abstracting massive data for lightweight intrusion detection in computer networks. Information Sciences **433-434**, 417–430 (2018). https://doi.org/10.1016/j.ins.2016.10.023

[15] Ambusaidi, M.A., He, X., Nanda, P., Tan, Z.: Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Transactions on Computers **65**(10), 2986–2998 (2016). https://doi.org/10.1109/TC.2016.2519914

[16] Lin, W.-C., Ke, S.-W., Tsai, C.-F.: Cann: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-Based Systems **78**, 13–21 (2015). https://doi.org/10.1016/j.knosys.2015.01.009

[17] Wang, W., He, Y., Liu, J., Gombault, S.: Constructing important features from massive network traffic for lightweight intrusion detection. IET Information Security **9**(6), 374–379 (2015)

[18] Marir, N., Wang, H., Feng, G., Li, B., Jia, M.: Distributed abnormal

behavior detection approach based on deep belief network and ensemble svm using spark. IEEE Access **6**, 59657–59671 (2018). https://doi.org/10.1109/ACCESS.2018.2875045

[19] Elmasry, W., Akbulut, A., Zaim, A.H.: Evolving deep learning architectures for network intrusion detection using a double pso metaheuristic. Computer Networks **168**, 107042 (2020). https://doi.org/10.1016/j.comnet.2019.107042

[20] Ashfaq, R.A.R., Wang, X.-Z., Huang, J.Z., Abbas, H., He, Y.-L.: Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences **378**, 484–497 (2017). https://doi.org/10.1016/j.ins.2016.04.019

[21] Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition **58**, 121–134 (2016). https://doi.org/10.1016/j.patcog.2016.03.028

[22] Chenghua, T., Xiang, Y., Wang, Y., Qian, J., Qiang, B.: Detection and classification of anomaly intrusion using hierarchy clustering and svm: Detection and classification of anomaly intrusion. Security and Communication Networks **9** (2016). https://doi.org/10.1002/sec.1547

[23] Alazzam, H., Sharieh, A., Sabri, K.E.: A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. Expert Systems with Applications **148**, 113249 (2020). https://doi.org/10.1016/j.eswa.2020.113249

[24] Hosseinzadeh Aghdam, M., Kabiri, P.: Feature selection for intrusion detection system using ant colony optimization. International Journal of Network Security **18**, 420–432 (2016)

[25] Gharib, A., Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: An evaluation framework for intrusion detection dataset. In: 2016 International Conference on Information Science and Security (ICISS), pp. 1–6 (2016)

[26] Gupta, M., Shrivastava, S.: Intrusion detection system based on svm and bee colony. International Journal of Computer Applications **111**, 27–32 (2015). https://doi.org/10.5120/19576-1377

[27] Feng, W., Zhang, Q., Hu, G., Huang, J.X.: Mining network data for intrusion detection through combining svms with ant colony networks. Future Generation Computer Systems **37**, 127–140 (2014). https://doi.org/10.1016/j.future.2013.06.027. Special Section: Innovative Methods and Algorithms for Advanced Data-Intensive Computing Special Section: Semantics, Intelligent processing and services for big data Special Section:

Advances in Data-Intensive Modelling and Simulation Special Section: Hybrid Intelligence for Growing Internet and its Applications

[28] Bamakan, S.M.H., Amiri, B., Mirzabagheri, M., Shi, Y.: A new intrusion detection approach using pso based multiple criteria linear programming. Procedia Computer Science **55**, 231–237 (2015). https://doi.org/10.1016/j.procs.2015.07.040. 3rd International Conference on Information Technology and Quantitative Management, ITQM 2015

[29] Othman, S.M., Ba-Alwi, F.M., Alsohybe, N.T., Al-Hashida, A.Y.: Intrusion detection model using machine learning algorithm on big data environment. Journal of Big Data **5**(1), 34 (2018). https://doi.org/10.1186/s40537-018-0145-4

[30] Kuang, F., Zhang, S., Jin, Z., Xu, W.: A novel svm by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. Soft Computing **19**(5), 1187–1199 (2015). https://doi.org/10.1007/s00500-014-1332-7

[31] la Hoz], E.D., Hoz], E.D.L., Ortiz, A., Ortega, J., Prieto, B.: Pca filtering and probabilistic som for network intrusion detection. Neurocomputing **164**, 71–81 (2015). https://doi.org/10.1016/j.neucom.2014.09.083

[32] Osanaiye, O., Cai, H., Choo, K.-K.R., Dehghantanha, A., Xu, Z., Dlodlo, M.: Ensemble-based multi-filter feature selection method for ddos detection in cloud computing. EURASIP Journal on Wireless Communications and Networking **2016**(1), 130 (2016). https://doi.org/10.1186/s13638-016-0623-3

[33] Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J., Alazab, A.: Hybrid intrusion detection system based on the stacking ensemble of c5 decision tree classifier and one class support vector machine. Electronics **9**, 173 (2020). https://doi.org/10.3390/electronics9010173

[34] Kunal, Dua, M.: Attribute selection and ensemble classifier based novel approach to intrusion detection system. Procedia Computer Science **167**, 2191–2199 (2020). https://doi.org/10.1016/j.procs.2020.03.271. International Conference on Computational Intelligence and Data Science

[35] Ji, S.-Y., Jeong, B.-K., Choi, S., Jeong, D.H.: A multi-level intrusion detection method for abnormal network behaviors. Journal of Network and Computer Applications **62**, 9–17 (2016). https://doi.org/10.1016/j.jnca.2015.12.004

[36] Choi, H., Kim, M., Lee, G., Kim, W.: Unsupervised learning approach for network intrusion detection system using autoencoders. The Journal of Supercomputing **75** (2019). https://doi.org/10.1007/

s11227-019-02805-w

[37] Kasongo, S.M., Sun, Y.: A deep long short-term memory based classifier for wireless intrusion detection system. ICT Express (2019). https://doi.org/10.1016/j.icte.2019.08.004

[38] Singh, R., Kumar, H., Singla, R.K.: An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Systems with Applications **42**(22), 8609–8624 (2015). https://doi.org/10.1016/j.eswa.2015.07.015

[39] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., Hotho, A.: A survey of network-based intrusion detection data sets (2019)

[40] Salo, F., Nassif, A.B., Essex, A.: Dimensionality reduction with ig-pca and ensemble classifier for network intrusion detection. Computer Networks **148**, 164–175 (2019). https://doi.org/10.1016/j.comnet.2018.11.010

[41] Jabbar, M.A., Aluvalu, R., S], S.S.R.: Rfaode: A novel ensemble intrusion detection system. Procedia Computer Science **115**, 226–234 (2017). https://doi.org/10.1016/j.procs.2017.09.129. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

[42] Aldwairi, T., Perera, D., Novotny, M.A.: An evaluation of the performance of restricted boltzmann machines as a model for anomaly network intrusion detection. Computer Networks **144**, 111–119 (2018). https://doi.org/10.1016/j.comnet.2018.07.025

[43] Tan, Z., Jamdagni, A., He, X., Nanda, P., Liu, R., Hu, J.: Detection of denial-of-service attacks based on computer vision techniques. IEEE Transactions on Computers **64** (2015). https://doi.org/10.1109/TC.2014.2375218

[44] Shiravi, A., Shiravi, H., Tavallaee, M., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security **31**(3), 357–374 (2012). https://doi.org/10.1016/j.cose.2011.12.012

[45] Ashfaq Khan, M., Karim, M., Kim, Y.: A scalable and hybrid intrusion detection system based on the convolutional-lstm network. Symmetry **11**, 583 (2019). https://doi.org/10.3390/sym11040583

[46] Moustafa, N., Slay, J.: 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), pp. 1–6 (2015)

[47] Lv, L., Wang, W., Zhang, Z., Liu, X.: A novel intrusion detection system based on an optimal hybrid kernel extreme learning machine. Knowledge-Based Systems **195**, 105648 (2020). https://doi.org/10.1016/j.knosys.2020.105648

[48] Adhi Tama, B., Comuzzi, M., Rhee, K.H.: Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. IEEE Access **7** (2019). https://doi.org/10.1109/ACCESS.2019.2928048

[49] Hajisalem, V., Babaie, S.: A hybrid intrusion detection system based on abc-afs algorithm for misuse and anomaly detection. Computer Networks **136**, 37–50 (2018). https://doi.org/10.1016/j.comnet.2018.02.028

[50] Ren, J., Guo, J., Qian, W., Yuan, H., Hao, X., u Jingjing: Building an effective intrusion detection system by using hybrid data optimization based on machine learning algorithms. Security and Communication Networks **2019**(4), 609–619 (2019). https://doi.org/10.1155/2019/7130868

[51] Zhang, J., Ling, Y., Fu, X., Yang, X., Xiong, G., Zhang, R.: Model of the intrusion detection system based on the integration of spatial-temporal features. Computers & Security **89**, 101681 (2020). https://doi.org/10.1016/j.cose.2019.101681

[52] Zhang, H., Lu, G., Qassrawi, M.T., Zhang, Y., Yu, X.: Feature selection for optimizing traffic classification. Computer Communications **35**(12), 1457–1471 (2012). https://doi.org/10.1016/j.comcom.2012.04.012

[53] Zhou, Y., Cheng, G., Jiang, S., Dai, M.: Building an efficient intrusion detection system based on feature selection and ensemble classifier. Computer Networks, 107247 (2020). https://doi.org/10.1016/j.comnet.2020.107247

[54] Kolias, C., Kambourakis, G., Stavrou, A., Gritzalis, S.: Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset. IEEE Communications Surveys & Tutorials **18**, 1–1 (2015). https://doi.org/10.1109/COMST.2015.2402161

[55] Creech, G., Hu, J.: A semantic approach to host-based intrusion detection systems using contiguousand discontiguous system call patterns. IEEE Transactions on Computers **63**(4), 807–819 (2014)

[56] Creech, G., Hu, J.: Generation of a new ids test dataset: Time to retire the kdd collection. In: 2013 IEEE Wireless Communications and Networking Conference (WCNC), pp. 4487–4492 (2013)

[57] Moore, A.W., Zuev, D., Crogan, M.L.: Discriminators for use in flow-based classification. (2013)

[58] Shi, H., Li, H., Zhang, D., Cheng, C., Cao, X.: An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. Computer Networks **132**, 81–98 (2018). https://doi.org/10.1016/j.comnet.2018.01.007

[59] Liu, Z., Wang, R., Tao, M., Cai, X.: A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion. Neurocomputing **168**, 365–381 (2015). https://doi.org/10.1016/j.neucom.2015.05.089

[60] tcptrace – Official Homepage. Available(2017.8.17) (2021). http://www.tcptrace.org/

[61] Gringoli, F., Salgarelli, L., Dusi, M., Cascarano, N., Risso, F., claffy, k.c.: Gt: Picking up the truth from the ground for internet traffic. SIGCOMM Comput. Commun. Rev. **39**(5), 12–18 (2009). https://doi.org/10.1145/1629607.1629610

[62] CSE-CIC-IDS2018 on AWS dataset (2021). https://www.unb.ca/cic/datasets/ids-2018.html

[63] Li, X., Chen, W., Zhang, Q., Wu, L.: Building auto-encoder intrusion detection system based on random forest feature selection. Computers & Security, 101851 (2020). https://doi.org/10.1016/j.cose.2020.101851