International Journal of Computers and Applications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjca20

A survey of intrusion detection from the perspective of intrusion datasets and machine learning techniques

Geeta Singh & Neelu Khare

INTERNATIONAL JOURNAL OF COMPUTERS AND APPLICATIONS
https://doi.org/10.1080/1206212X.2021.1885150

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A survey of intrusion detection from the perspective of intrusion datasets and machine learning techniques

Geeta Singh[a] and Neelu Khare[b]

[a]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India; [b]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

ABSTRACT

The evolution in the attack scenarios has been such that finding efficient and optimal Network Intrusion Detection Systems (NIDS) with frequent updates has become a big challenge. NIDS implementation using machine learning (ML) techniques and updated intrusion datasets is one of the solutions for effective modeling of NIDS. This article presents a brief description of publicly available labeled intrusion datasets and ML techniques. Later a brief explanation of the literary works is given in which machine learning techniques are applied for NIDS implementation in different networking scenarios, such as traditional networks, cloud networks, Ad-Hoc, WSNs, and IoT networks. Hence, this article brings together publicly available intrusion datasets and machine learning techniques utilized in recent intrusion detection systems to reveal present-day challenges and future directions. This article also explains problems associated with NIDS. This will help researchers to enhance the existing NIDS models as well as to develop new effective models.

ARTICLE HISTORY
Received 24 July 2020
Accepted 30 January 2021

KEYWORDS
Intrusion detection; anomaly; intrusion dataset; machine learning; security

1. Introduction

In recent years, malicious activities such as illegal data access, stolen credential, impersonation, data alteration, intrusion, and many more are spreading all over the cyber world at an alarming rate. Existing security and prevention methods do not remain sufficient to provide complete protection from sophisticated attacks and malware which are evolving and making the detection process more complicated and challenging. Security requirements of Cloud network, Wireless Sensor Network (WSN), Ad-Hoc networks, and IoT networks are different from the traditional network. Network Intrusion Detection System (NIDS) automates the detection of intrusion activities that can compromise the confidentiality, availability, and integrity of a computer system or computer network through bypassing the security mechanisms. Hence, improvement in NIDS is an urgent necessity. Intrusion detection approaches are mainly categorized into three categories: (1) Signature-based IDS stores the signature of known attack types and uses this signature base to find similar attacks in the network traffic (2) Anomaly-based IDS detects anomalies in system behavior by matching it with stored normal system behavior (3) Hybrid IDS employs both the approaches together. False Positive (FP) and False-Negative (FN) are the states when an IDS wrongly identifies normal behavior as attack and attack behavior as normal behavior respectively [1–3]. Intrusion detection models are usually evaluated using the intrusion datasets that contain normal and anomalous network traffic patterns. Old intrusion datasets have now become outdated because of evolving attack tactics. These datasets must contain modern attacks as well as genuine network traffics to enhance the detection accuracy of NIDS. The adaptability of ML techniques to a new environment is a useful characteristic for security applications. ML techniques automate the process of attack detection and aid in efficiently building NIDS models. ML-based NIDS fulfills the current security needs to some extent. According to Gartner's survey [4], cybersecurity is one of the topmost technological investment areas in 2019, and Artificial Intelligence (AI)/ML is above all game-changing technologies for government Chief Information Officers (CIOs) for 2019 (see Figure 1). These facts motivated us to explore new approaches to computer security applying AI/ ML techniques.

Computer networks are vulnerable to attacks if do not have a security plan in an appropriate place [5]. Hence the importance of this paper is multifold as this paper focuses on the following points.

(1) This paper discusses some popular and latest ML algorithms to reveal their characteristics and limitations. This will help the researchers to select an appropriate algorithm for carrying out their research.
(2) It describes commonly used intrusion datasets. Periodic assessment of intrusion datasets plays a vital role in attaining NIDS goals. It helps in selecting the appropriate dataset for the evaluation of a specific NIDS. This also helps in dataset enhancement.
(3) A periodic assessment of existing intrusion detection models is necessary to reveal the recent advancement and challenges in NIDS modeling. A deep analysis of different network domains is performed to unfold their security concerns and limitations that will aid in enhancing the performance of existing NIDS and implementing new improved models.
(4) FP, FN, data imbalance, etc., are common problems that degrade NIDS performance. These problems are discussed to attract the attention of the researchers so that they can gain important insights into how to improve NIDS performance.

This paper has four main contributions.

- 23 existing works are considered to describe the publicly available labeled intrusion datasets. Data source, traffic category,

CONTACT Neelu Khare neelu.khare@vit.ac.in, neelukh.29@gmail.com

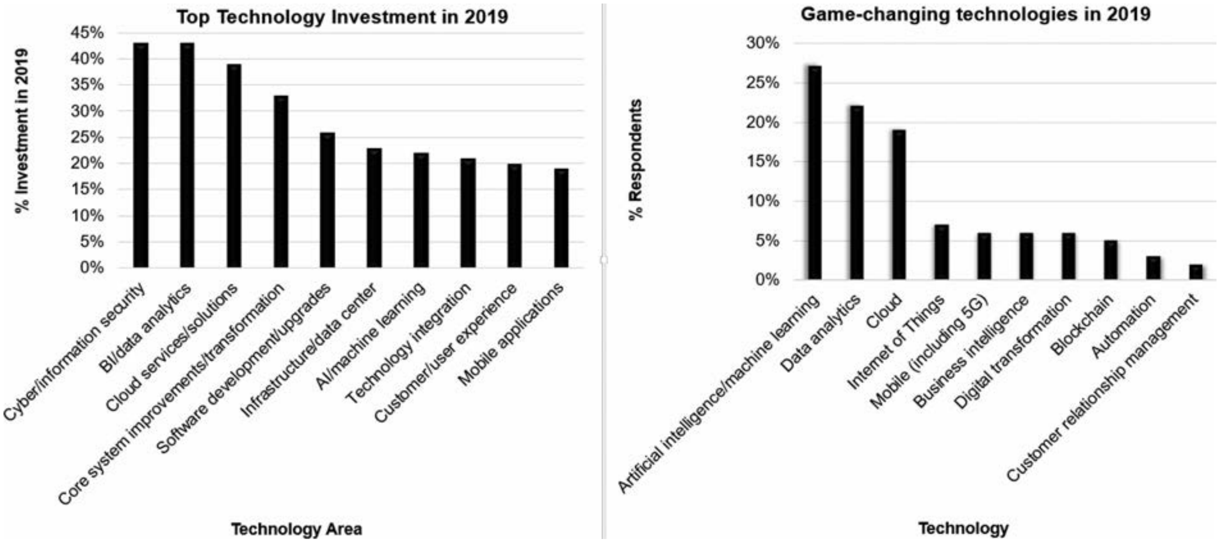


Figure 1. Gartner's survey 2019 statistics.

characteristics, demerits, features, and attack types of the data sets are summarized in a tabular format.

- 26 existing works are considered to analyze some popular ML algorithms to get an insight into their characteristics, uses, and limitations in the intrusion detection process.
- 23 recent ML-based NIDS models built using public intrusion datasets are surveyed to highlight the current challenges, proposed solutions, results, and future directions
- Finally, this article draws attention to the problems that vitiate the process of intrusion detection.

This article is organized as follows: A brief description and comparison of various intrusion datasets are given in section 2. Section 3 describes some popular and some latest ML techniques that have been applied for IDS implementation. Recent IDS models are surveyed in section 4. Section 5 focuses on the problems associated with NIDS. The conclusion is given in Section 6.

2. Publicly available intrusion datasets

Intrusion datasets are surveyed and analyzed in several existing literature works with different objectives [7–9]. This section describes publicly available intrusion datasets namely KDD Cup '99, Network Security Laboratory-KDD (NSL-KDD), Aegean Wi-Fi Intrusion Dataset (AWID), Yahoo Webscope S5 anomaly benchmark, Numenta Anomaly Benchmark (NAB), Kyoto 2006+, UNSW-NB 15, BoT_IoT, Drebin, Contagio, and Genome. To the best of our knowledge, such a variety of intrusion datasets is not discussed in any existing survey.

2.1. KDD Cup '99

In 1998, the Defense Advanced Research Projects Agency (DARPA), a division of the U.S. Defense Unit conducted an evaluation program in MIT Lincoln Labs to serve the objective of examining intrusion detection researches [10]. An extensive range of intrusion attack traffic was simulated in the U.S. Air Force LAN environment. KDD cup 99 comprises a set of these traffics [11]. KDD Cup '99 dataset has a total of 41 attributes and one more field namely attack class that labels all the observations into normal or the attacks that fall under one of the four categories: Denial-of-Service (DoS), Remote to Local (R2L), User to Root (U2R), and Probing/surveillance. KDD Cup '99 dataset resulted in biased classification due to some inherent flaws in the dataset such as redundancy and missing values in observations [12].

2.2. NSL-KDD dataset

It is an improvement over the KDD Cup '99 dataset in which

- Insignificant observations are removed from the training dataset. This resulted in unbiased classifier generation towards the more frequent record.
- Duplicate records are removed in the test sets. This resulted in unbiased learners' performance towards the approaches that otherwise better classify frequent records only.
- KDD dataset records have a different degree of difficulty. For the preparation of NSL-KDD, records are selected in inverse proportion to the percentage of records in the whole dataset. This resulted in efficient evaluation accuracy of diverse learning methods.
- A fair quantity of train and test dataset records in NSL-KDD leads to the efficient execution of experiments, Consistent and comparable evaluation results.

The attack labels count and the attributes count in the NSL-KDD dataset are similar to those in the KDD Cup '99 dataset [13]. But this new version also suffers from the problems discussed in [14]. NSL-KDD doesn't provide exact definitions of the attacks and doesn't represent existing real networks. Its compatibility with real network traffic is not verified. Despite the flaws, KDD Cup '99 and NSL-KDD datasets are still being used in many recent intrusion detection research works [15,16].

2.3. AWID dataset

AWID Offers tools, methodologies to implement wireless network IDS. AWID dataset comprises WLAN traffic in packet-based format. AWID comprises two versions: Large dataset and Reduced dataset, which are further subdivided into a high-level labeled dataset and Finer grained labeled dataset. This dataset has 155 features, including the class label [17]. AWID is imbalanced. So it needs proper pre-processing before use.

2.4. Yahoo Webscope s5

Yahoo Webscope S5 dataset is generated by the Media Sciences team at Yahoo Labs to detect anomalies in time series data [18]. It has a goal of benchmarking the anomaly detection algorithm. It is a blend of synthetic time-series and real Yahoo services traffic. The simulated dataset has algorithmically generated anomalies. On the other hand, the real-traffic dataset anomalies are manually labeled. This dataset comprises four different classes namely A1, A2, A3, and A4 with tagged time-series counts 67, 100, 100, and 100 respectively. These four classes comprise 367 time series of length 1500. Class A1 contains real data while the remaining three Classes contain artificial anomaly data. These datasets are growing over time by adding more time-series data.

2.5. NAB dataset

It is a standard benchmark to detect anomalies in streaming data [19]. This dataset includes two types of data: ***Real Data***-realAWSCloudwatch, realAdExchange, realKnownCause, realTraffic, realTweets, and ***Artificial Data***-artificialNoAnomaly, artificialWithAnomaly. NAB version 1.0 consists of 58 data streams. All the streams have around 1000–22,000 records representing 365,551 data points. It includes 11 simulated data files with anomalous behaviors. It also includes some additional data files with normal behavior. All data files are either labeled with known anomalies or with a well-defined NAB labeling procedure. Data in this dataset are ordered, time-stamped, and single-valued metrics. NAB labeled dataset has some inherent challenges [20]. Some of the NAB datasets have an unequal interval of observation times. Due to this reason, these datasets are not compatible with ML as well as traditional time series frameworks. The difference in data distribution leads to a variance in training and test partition that further causes low performance of different models.

2.6. Kyoto 2006+

Kyoto 2006+ is a real traffic dataset acquired from different honeypots [21,22]. It has 14 conventional features and 10 additional features. Some unknown IDS alerts, AV alerts, and shellcodes are observed in honeypot data. Kyoto dataset has 13 Connection States and three undefined class labels; Normal, Abnormal, and Unknown. This dataset is troublesome to download and run the experiments on it. It has a significantly low amount of features as well as the quantity and range of real ordinary user activities is also less.

2.7. UNSW NB-15 dataset

UNSW NB-15 dataset comprises real recent normal network traffic traces as well as recent synthesized anomalous traffic activities [23,24]. This dataset includes a total of 2,540,038 flows out of which 2,218,755 are legitimate flows and 321,283 are attack flows. It has a total of 49 features including class labels. Many additional features are suitable for the detection of new types of attacks. Contemporary low footprint attacks are eventually reflected by the attack groups.

2.8. BoT_IoT dataset

BoT_IoT dataset includes real and simulated IoT network traffic [25]. The traffic comprises ordinary traffic and botnet traffic with 73,370,443 records. The BoT_IoT dataset has a total of 46 features including three class labels namely ‘attack,’ ‘category,’ and ‘subcategory.’ The ‘attack’ class label has two values; ‘0’ for normal traffic and ‘1’ for attack traffic. The ‘category’ class label divides the attack traffics into 3 categories which are further subdivided into 6 subcategories by the ‘subcategory’ class label.

2.9. Malware datasets- Drebin, Contagio, and Genome

Drebin, Contagio, and Genome are popular malware datasets, which are used to detect and classify widespread malware [26]. Genome dataset contains different types of Android malware (Collection duration: August 2010- October 2011) [27]. Drebin dataset contains real Android malware and real Android application samples from different websites (Collection duration: August 2010- October 2012) [28]. The Contagio dataset contains mobile malware samples as well as benign samples. This dataset is online accessible at Contagio Malware Dump [29].

Figure 2 describes different intrusion datasets in terms of source, traffic category, merits, demerits, features. and anomalies.

The behavior and limitations of the ML method must be known before applying it for NIDS modeling using any particular dataset. The next section discusses some ML methods popular in NIDS modeling.

3. Machine learning techniques

A variety of ML algorithms have already been exploited to implement NIDS models for different purposes [6]. These ML algorithms can be broadly divided into five categories based on learning strategies. ML-algorithm categorization, description, and examples are shown in Figure 3. This section mainly focuses on the characteristics and limitations of popular ML algorithms.

3.1. Decision tree

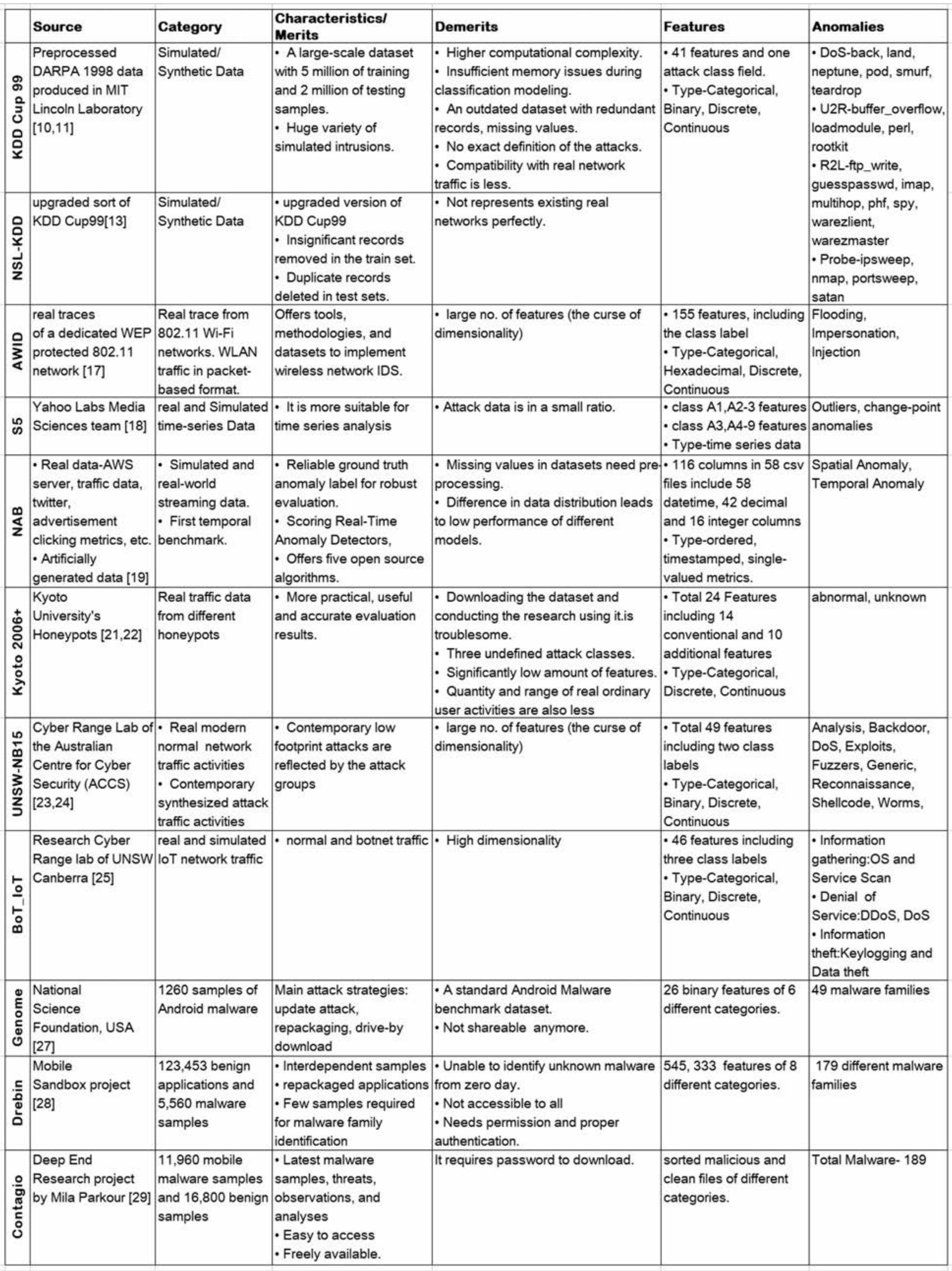
DT is a hierarchical representation of all the possible solutions to a decision making or classification problem [31]. It classifies given problem space into predefined classes following a sequence of tests for each attribute value. These tests are conducted by the internal nodes of the decision tree. Branches represent possible outcomes of the test. External nodes represent the classes of the partitioned problem spaces. DT is easy to build, use, and interpret. Despite these benefits, trees are not an ideal technique for predictive learning. DT provides good accuracy with the known dataset. But the model may not classify the unknown dataset accurately. This is due to overfitting. Thus Miss Rate is more as compared to some other algorithms.

3.2. Random forest

RF is one of the ensemble classifiers which combines the predicted results from different decision tree predictors to output final prediction [32]. The decision of the majority of the tree is chosen by the RF as the final decision. Many benefits are associated with RF that compensate for the limitations and drawbacks of the decision tree. RF is one solution for overfitting problems associated with decision trees [33]. It can handle large databases with higher dimensionality and can run efficiently. RF predictions are highly accurate for large datasets, even though a large portion of the dataset contains missing values.

3.3. Deep neural networks

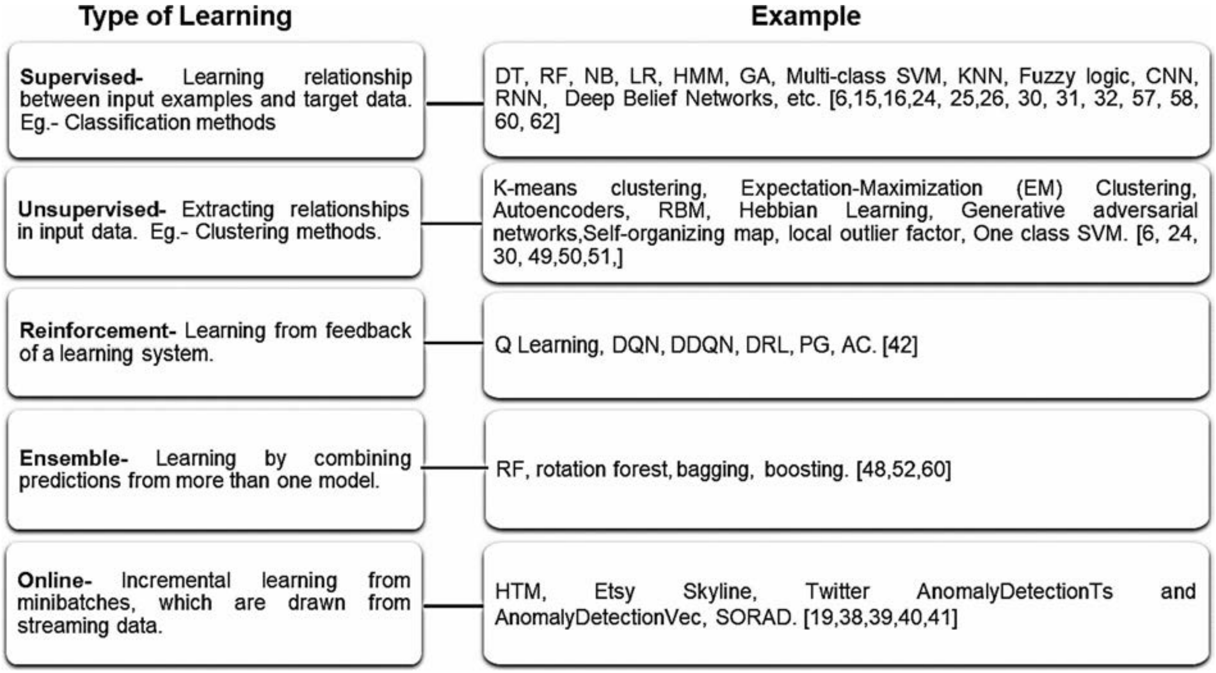
DNNs are artificial neural networks (ANN) that have various layers in the middle of the input and output layers [34]. Neuron layers are stacked along the depth and width to make a network architecture which decreases the number of features at each stack and converts the representation into a more abstract arrangement.



	Source	Category	Characteristics/ Merits	Demerits	Features	Anomalies
KDD Cup 99	Preprocessed DARPA 1998 data produced in MIT Lincoln Laboratory [10,11]	Simulated/ Synthetic Data	• A large-scale dataset with 5 million of training and 2 million of testing samples. • Huge variety of simulated intrusions.	• Higher computational complexity. • Insufficient memory issues during classification modeling. • An outdated dataset with redundant records, missing values. • No exact definition of the attacks. • Compatibility with real network traffic is less.	• 41 features and one attack class field. • Type-Categorical, Binary, Discrete, Continuous	• DoS-back, land, neptune, pod, smurf, teardrop • U2R-buffer_overflow, loadmodule, perl, rootkit • R2L-ftp_write, guesspasswd, imap, multihop, phf, spy, warezlient, warezmaster • Probe-ipsweep, nmap, portsweep, satan
NSL-KDD	upgraded sort of KDD Cup99[13]	Simulated/ Synthetic Data	• upgraded version of KDD Cup99 • Insignificant records removed in the train set. • Duplicate records deleted in test sets.	• Not represents existing real networks perfectly.		
AWID	real traces of a dedicated WEP protected 802.11 network [17]	Real trace from 802.11 Wi-Fi networks. WLAN traffic in packet-based format.	Offers tools, methodologies, and datasets to implement wireless network IDS.	• large no. of features (the curse of dimensionality)	• 155 features, including the class label • Type-Categorical, Hexadecimal, Discrete, Continuous	Flooding, Impersonation, Injection
S5	Yahoo Labs Media Sciences team [18]	real and Simulated time-series Data	• It is more suitable for time series analysis	• Attack data is in a small ratio.	• class A1,A2-3 features • class A3,A4-9 features • Type-time series data	Outliers, change-point anomalies
NAB	• Real data-AWS server, traffic data, twitter, advertisement clicking metrics, etc. • Artificially generated data [19]	• Simulated and real-world streaming data. • First temporal benchmark.	• Reliable ground truth anomaly label for robust evaluation. • Scoring Real-Time Anomaly Detectors, • Offers five open source algorithms.	• Missing values in datasets need pre-processing. • Difference in data distribution leads to low performance of different models.	• 116 columns in 58 csv files include 58 datetime, 42 decimal and 16 integer columns • Type-ordered, timestamped, single-valued metrics.	Spatial Anomaly, Temporal Anomaly
Kyoto 2006+	Kyoto University's Honeypots [21,22]	Real traffic data from different honeypots	• More practical, useful and accurate evaluation results.	• Downloading the dataset and conducting the research using it.is troublesome. • Three undefined attack classes. • Significantly low amount of features. • Quantity and range of real ordinary user activities are also less	• Total 24 Features including 14 conventional and 10 additional features • Type-Categorical, Discrete, Continuous	abnormal, unknown
UNSW-NB15	Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) [23,24]	• Real modern normal network traffic activities • Contemporary synthesized attack traffic activities	• Contemporary low footprint attacks are reflected by the attack groups	• large no. of features (the curse of dimensionality)	• Total 49 features including two class labels • Type-Categorical, Binary, Discrete, Continuous	Analysis, Backdoor, DoS, Exploits, Fuzzers, Generic, Reconnaissance, Shellcode, Worms,
BoT_IoT	Research Cyber Range lab of UNSW Canberra [25]	real and simulated IoT network traffic	• normal and botnet traffic	• High dimensionality	• 46 features including three class labels • Type-Categorical, Binary, Discrete, Continuous	• Information gathering:OS and Service Scan • Denial of Service:DDoS, DoS • Information theft:Keylogging and Data theft
Genome	National Science Foundation, USA [27]	1260 samples of Android malware	Main attack strategies: update attack, repackaging, drive-by download	• A standard Android Malware benchmark dataset. • Not shareable anymore.	26 binary features of 6 different categories.	49 malware families
Drebin	Mobile Sandbox project [28]	123,453 benign applications and 5,560 malware samples	• Interdependent samples • repackaged applications • Few samples required for malware family identification	• Unable to identify unknown malware from zero day. • Not accessible to all • Needs permission and proper authentication.	545, 333 features of 8 different categories.	179 different malware families
Contagio	Deep End Research project by Mila Parkour [29]	11,960 mobile malware samples and 16,800 benign samples	• Latest malware samples, threats, observations, and analyses • Easy to access • Freely available.	It requires password to download.	sorted malicious and clean files of different categories.	Total Malware- 189

Figure 2. Comparison of intrusion datasets.

Applying neural networks for intrusion detection has many benefits [35]. But, it needs voluminous input to outperform the other algorithms. Moreover, training a model with DNN is expensive in terms of resource requirements. The selection of an appropriate DNN candidate requires deep knowledge of the underlying architecture, methods, and parameters.



Type of Learning	Example
Supervised- Learning relationship between input examples and target data. Eg.- Classification methods	DT, RF, NB, LR, HMM, GA, Multi-class SVM, KNN, Fuzzy logic, CNN, RNN, Deep Belief Networks, etc. [6,15,16,24, 25,26, 30, 31, 32, 57, 58, 60, 62]
Unsupervised- Extracting relationships in input data. Eg.- Clustering methods.	K-means clustering, Expectation-Maximization (EM) Clustering, Autoencoders, RBM, Hebbian Learning, Generative adversarial networks,Self-organizing map, local outlier factor, One class SVM. [6, 24, 30, 49,50,51,]
Reinforcement- Learning from feedback of a learning system.	Q Learning, DQN, DDQN, DRL, PG, AC. [42]
Ensemble- Learning by combining predictions from more than one model.	RF, rotation forest, bagging, boosting. [48,52,60]
Online- Incremental learning from minibatches, which are drawn from streaming data.	HTM, Etsy Skyline, Twitter AnomalyDetectionTs and AnomalyDetectionVec, SORAD. [19,38,39,40,41]

Figure 3. Types of machine learning methods and examples.

3.4. Support vector machine

SVM converts the data in higher dimensions to find the hyper-plane that separates the input data between the anomalous and normal data instances with the maximal margin [36]. And thus it minimizes the upper bound on the generalization error. SVM shows limited performance for learning problems while applied to imbalanced datasets. In some cases, negative instances might be more than positive instances [37]. Moreover, SVM is inapt for processing huge datasets with a large number of instances.

3.5. Hierarchical temporal memory

In humans, the neocortex performs complex tasks such as object identification and manipulation by touch, visual pattern recognition, spoken language comprehension, etc. HTM technology implements human-level cognitive tasks performed by the neocortex [38]. HTM is a time-based continuous learning algorithm similar to the neural network. HTM store and recall spatial and temporal patterns. It scores anomalies between 0 and 1. Mean and variance of the latest anomaly scores distribution is maintained by the system to correlate the current anomaly score with respective normal distribution at every time step. The final anomaly is predicted by thresholding the anomaly likelihood. The Numenta Platform for Intelligent Computing (NuPIC) provides a platform to implement the HTM learning algorithms. HTM attains significant performance improvements on NAB benchmarks. But it does not perform so well on the Yahoo Webscope S5 dataset.

3.6. Twitter ADVec

Twitter ADVec is developed by Twitter [39]. It is a composition of diverse algorithms. The primary algorithm, Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) is an anomaly detection approach to detect global and local abnormalities in time-series data. This approach is based on the Generalized ESD. This algorithm proceeds by first decomposing time series data and then calculating median and ESD. Piecewise approximation is also employed in this approach for long time-series data.

3.7. Skyline

It is a real-time robust anomaly detection scheme that facilitates the passive monitoring of numerous metrics [40]. It is an appropriate choice for streaming data analysis. It can constantly monitor a large quantity of high-resolution time series. It is an easily extendible algorithm that automatically detects an anomalous metric and treats the entire time series. It provides two services:

- Horizon agent- It takes care of new data points and keeps the time series sanitized and updated.
- Analyzer agent- Responsible for analyzing every metric for anomalies.

In this system, a set of simple detectors are applied to detect nonconformity from a least-squares evaluation, past values histogram, moving average, and so on. A voting system gives the concluding anomaly score. Open source code for the skyline is available to use.

3.8. Simple online regression anomaly detector (SORAD).

It is an Anomaly detection algorithm for time series data that requires refinement, extension, and test on more varied anomaly data benchmarks [41]. SORAD has several parameters as forgetting parameter, anomaly threshold, and windows-size to set before applying the algorithm. It has a short-term memory. Due to this reason, recurring spikes in the time series are identified as an anomaly for which SORAD generates many False Positives. SORAD examines a short-term window in the original time series and therefore it is not suitable for long-range anomaly detection.

3.9. Reinforcement learning

In Reinforcement learning, a model is trained to take action sequentially in a complex unfamiliar environment to attain a particular goal set by the programmer. It is a trial and error based solution. Each action is either get a reward or a penalty. The overall aim is to achieve higher rewards [42].

The next section surveys the NIDS models build using a variety of ML techniques including all the techniques discussed in this section.

4. NIDS in diverse networks

This section reviews ML-based NIDS models deployed in traditional as well as novel networking environments. KDD Cup '99 and NSL-KDD are the two most widely used Intrusion datasets in the domain of NIDS. But due to some flaws, alternatives to these datasets are explored in literature. UNSW-NB 15 and BoT_IoT is found as a substitute for these two datasets to classify contemporary attacks. This section is limited to review NIDS models experimented on one of these four datasets (KDD Cup '99, NSL-KDD, UNSW-NB 15, and BoT_IoT), or a combination of these datasets. This section discusses the challenges and suggestions about IDS modeling with the findings of the existing IDS models developed in a variety of network infrastructures. Future directions are suggested by the discussed existing models that will draw the attention of the researchers in the right direction.

4.1. NIDS in traditional networks

Model overfitting and training time complexity are big challenges in building efficient NIDS. A hybrid IDS model XGBoost–DNN is proposed for binary classification of the NSL-KDD dataset [15]. XGBoost based feature importance score is applied to select relevant features for classification. Adam optimizer is used in the proposed method to optimize the DNN classifier. Fast and platform-independent NIDS is the outcome of this method. Further modification in this model is expected to attack multi-classification purposes.

Proper treatment of poor quality and high dimension intrusion dataset is necessary for ML-based IDS implementation. To achieve the goals, NIA based optimization technique, namely Spider Monkey Optimization (SMO) is employed for dimension reduction of the NSL-KDD dataset [16]. This binary classification approach employed DNN as a binary classifier. This approach significantly reduced DNN classifier training time and enhanced overall classification results. Extension of the proposed model for multiclass anomaly detection in IoT networks and malware detection in cloud and fog networks are future goals to be achieved. The adoption of the Nature-Inspired Algorithm (NIA) is attracting a lot of attention [43–47].

Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Genetic Algorithm (GA) are hybridized for feature selection [48]. Attack classification is accomplished in two-stages applying two meta classifiers: (1) Rotation Forest (2) Bagging. The binary classifier model achieves improved accuracy, specificity, and precision. This model can be extrapolated to multi-classification in the future.

Support Vector Machine (SVM) is a very popular ML-based algorithm adopted as a classifier in numerous attack classification procedures. Authors in [49] drew attention to improve the detection ability of SVM based IDS. The model employs Random Under Sampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE) for data balancing. t-distributed stochastic neighbor embedding (t-SNE) technique is used for dimension reduction. This SVM based model Boosts the detection rate for almost all the attack types. In the future, further improvement is expected in the U2R attacks detection rate by exploiting similarity measures instead of distance measures in t-SNE.

Exponentially increasing unknown threats are difficult to detect as these novel threats have low footprints for exploitation in the attack detection process. A soft-max classifier and stacked autoencoder-based Two-Stage Deep Learning (TSDL) model is proposed for real-time intrusion detection [50]. This model achieved high detection rates for novel attacks. In the future, this efficient model can be further optimized with reinforcement [42] and multi-task learning algorithms.

A few available records of new attacks lead to the generation of imbalanced data. IDS models trained on such kind of imbalanced data usually demonstrate low detection rates and high false alarms. Yang et al. [51] experimented with two imbalanced datasets namely NSL-KDD and UNSW-NB15 to implement an IDS. ICVAE encoder is used for automatic dimensionality reduction and DNN classifier weight initialization. This model achieved improved overall accuracy, recall, and false-positive rate, the higher detection rate of rare attacks, and unknown attacks. This model can be further extended to scrutinization of ICVAE latent variable through adversarial learning methods to better synthesize minority classes.

Iwendi et al. [52], aimed at improvement in FAR, accuracy, and detection rate of NIDS. The authors proposed an ensemble model for binary classification and multiclass classification of KDD99 and NSLKDD datasets. These datasets are first reduced with a correlation-based feature selection (CFS) approach and then classified with an ensemble of Classifiers using AdaBoost and Bagging techniques. J48, Reptree, and RF algorithms are exploited as the base classifiers. This model produces a low FAR and a high detection rate. The majority of classes show high classification accuracy but some minority classes show zero classification accuracy. This model can be further improved by applying appropriate data resampling methods.

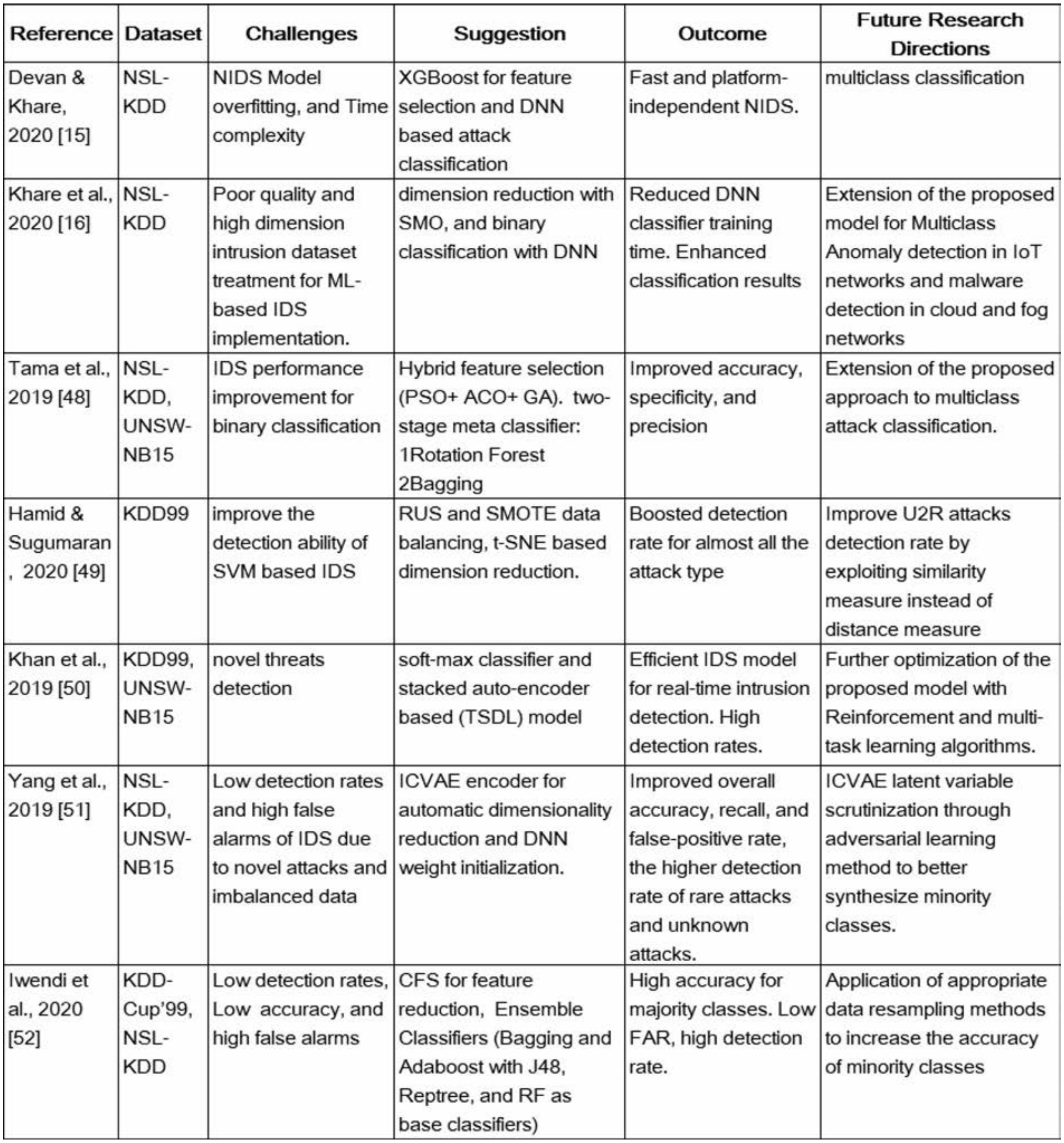
Figure 4 summarizes the above mentioned NIDS models.

4.2. NIDS in cloud, MANET and WANET, WSN and IoT networks

Huge networks like cloud environments and large IT ecosystems need a collaborative and extremely well-organized technique for attaining their security objectives. For the protection of such networks, collaborative Intrusion Detection Systems (CIDS) have emerged to detect sophisticated and highly distributed attacks [53]. Mobile Ad-Hoc Networks (MANETs) have inherent vulnerability features like open medium, highly dynamic network topology, limited physical security, lack of centralized monitoring and control system, etc. [54]. Vehicular ad hoc network (VANET) technology is grabbing the attention of all modern transportation systems. This technology is also vulnerable to different kinds of attacks [55].

VANET nodes can share their experiences and thus they can improve attack detection accuracy. Distributed ML is an appropriate structure for the implementation of this kind of cooperative attack/anomaly detection over VANETs. This Collaborative learning is also prone to attack as a malicious node can infer sensitive information from the data shared by other nodes in the network. The privacy-preserving ML-based collaborative IDS (PML-CIDS) algorithm can be used as a classifier to detect the intrusion type [56]. This algorithm's privacy notation is captured by differential privacy methods. Here, training data privacy protection and optimized security and privacy in VANET are the main objectives. NSL-KDD dataset is used in this work.

Bigdata techniques are also adopted in VANETs for handling a huge volume of data. Spark-ML RF-Based detection algorithm is suggested for DDoS attack detection in Bigdata generated from VANET [57]. A Micro-batch data processing technique is used for network traffic collection and feature extraction. Experiments are conducted on NSL-KDD and UNSW-NB15 datasets. Better accuracy and false positive rate (FPR) are the achievements of this experiment. The author suggested the deployment of the proposed NIDS in a real environment as future work.



Reference	Dataset	Challenges	Suggestion	Outcome	Future Research Directions
Devan & Khare, 2020 [15]	NSL-KDD	NIDS Model overfitting, and Time complexity	XGBoost for feature selection and DNN based attack classification	Fast and platform-independent NIDS.	multiclass classification
Khare et al., 2020 [16]	NSL-KDD	Poor quality and high dimension intrusion dataset treatment for ML-based IDS implementation.	dimension reduction with SMO, and binary classification with DNN	Reduced DNN classifier training time. Enhanced classification results	Extension of the proposed model for Multiclass Anomaly detection in IoT networks and malware detection in cloud and fog networks
Tama et al., 2019 [48]	NSL-KDD, UNSW-NB15	IDS performance improvement for binary classification	Hybrid feature selection (PSO+ ACO+ GA). two-stage meta classifier: 1Rotation Forest 2Bagging	Improved accuracy, specificity, and precision	Extension of the proposed approach to multiclass attack classification.
Hamid & Sugumaran , 2020 [49]	KDD99	improve the detection ability of SVM based IDS	RUS and SMOTE data balancing, t-SNE based dimension reduction.	Boosted detection rate for almost all the attack type	Improve U2R attacks detection rate by exploiting similarity measure instead of distance measure
Khan et al., 2019 [50]	KDD99, UNSW-NB15	novel threats detection	soft-max classifier and stacked auto-encoder based (TSDL) model	Efficient IDS model for real-time intrusion detection. High detection rates.	Further optimization of the proposed model with Reinforcement and multi-task learning algorithms.
Yang et al., 2019 [51]	NSL-KDD, UNSW-NB15	Low detection rates and high false alarms of IDS due to novel attacks and imbalanced data	ICVAE encoder for automatic dimensionality reduction and DNN weight initialization.	Improved overall accuracy, recall, and false-positive rate, the higher detection rate of rare attacks and unknown attacks.	ICVAE latent variable scrutinization through adversarial learning method to better synthesize minority classes.
Iwendi et al., 2020 [52]	KDD-Cup'99, NSL-KDD	Low detection rates, Low accuracy, and high false alarms	CFS for feature reduction, Ensemble Classifiers (Bagging and Adaboost with J48, Reptree, and RF as base classifiers)	High accuracy for majority classes. Low FAR, high detection rate.	Application of appropriate data resampling methods to increase the accuracy of minority classes

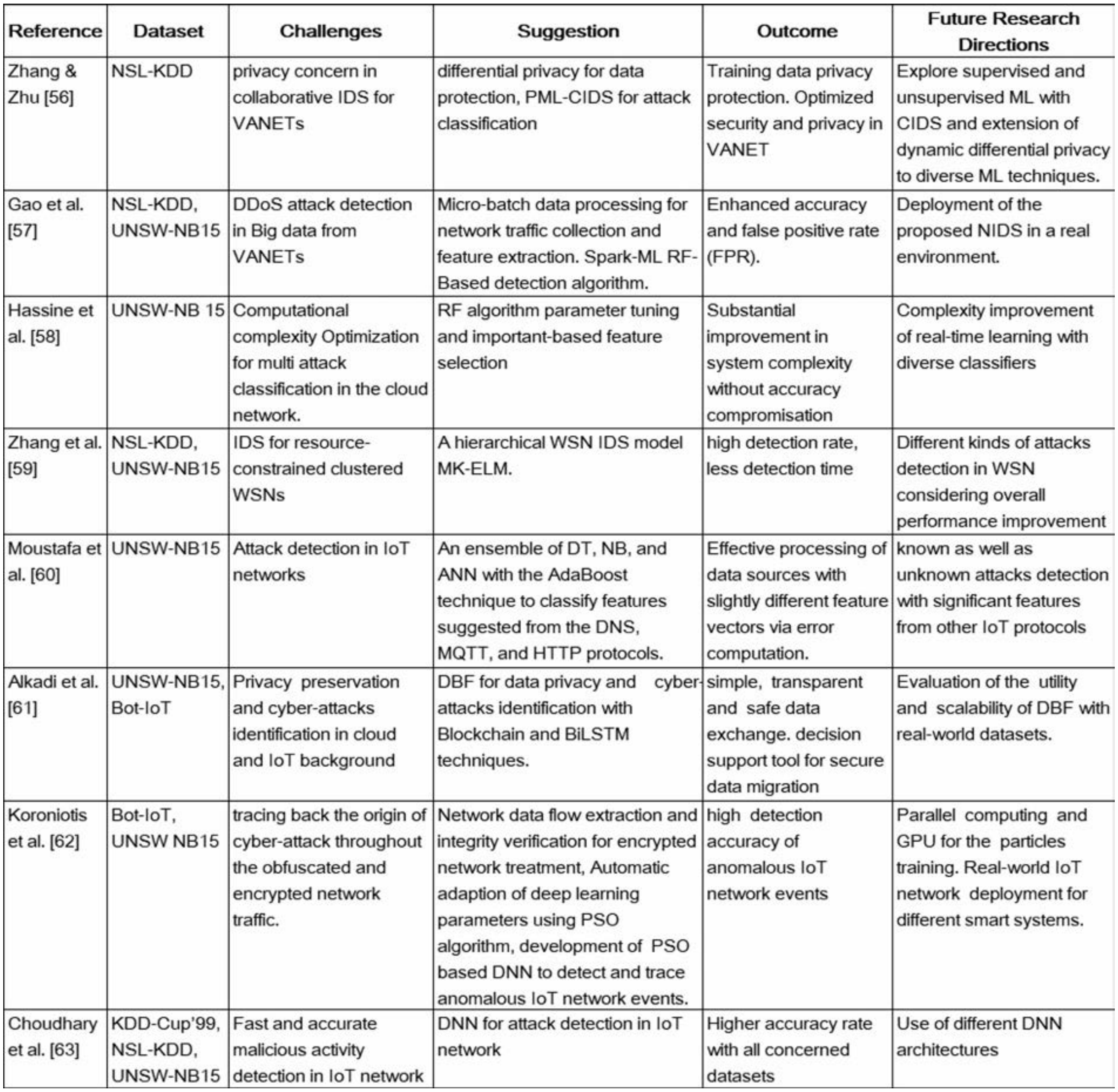
Figure 4. NIDS in traditional networks.

Nowadays, researchers are more focused on optimized intrusion detection solutions. Computational complexity Optimization has been a hot topic for cloud networks. Hassine et al. [58] proposed a multi-attack classification method in the cloud network. RF algorithm parameters are tuned to shrink the forest during the classification process while ensuring good accuracy. The importance-based feature selection method of the RF algorithm is used to find the optimal feature set. This approach attains a substantial improvement in system complexity without compromising accuracy and can be extended to complexity improvement of real-time learning with diverse classifiers.

Open deployment area, broadcast facility, and increased network connectivity of WSNs are some of the factors that expose the entire WSN infrastructure to external threats. Deployment of IDS is a necessity in WSN. A hierarchical WSN IDS model based on a multi-kernel extreme learning machine (MK-ELM) is proposed for resource-constrained clustered WSNs [59]. This model attains a high detection rate and less detection time. This work has the future scope considering less consumption of network energy and improved overall performance of different kinds of attack detection in WSN.

An ensemble of DT, NB, and ANN with the AdaBoost technique is proposed for attack detection in IoT networks [60]. Features are suggested from the DNS, MQTT, and HTTP protocols. The experiment is aimed at the effective processing of data sources with slightly different feature vectors via error computation. Other IoT protocols are also expected to be considered in future work for extracting significant features to detect known as well as unknown attacks.

Privacy preservation and cyber-attacks identification in cloud and IoT background is a challenging task. A deep Blockchain Framework (DBF) enforces security and privacy in the cloud and IoT



Reference	Dataset	Challenges	Suggestion	Outcome	Future Research Directions
Zhang & Zhu [56]	NSL-KDD	privacy concern in collaborative IDS for VANETs	differential privacy for data protection, PML-CIDS for attack classification	Training data privacy protection. Optimized security and privacy in VANET	Explore supervised and unsupervised ML with CIDS and extension of dynamic differential privacy to diverse ML techniques.
Gao et al. [57]	NSL-KDD, UNSW-NB15	DDoS attack detection in Big data from VANETs	Micro-batch data processing for network traffic collection and feature extraction. Spark-ML RF-Based detection algorithm.	Enhanced accuracy and false positive rate (FPR).	Deployment of the proposed NIDS in a real environment.
Hassine et al. [58]	UNSW-NB 15	Computational complexity Optimization for multi attack classification in the cloud network.	RF algorithm parameter tuning and important-based feature selection	Substantial improvement in system complexity without accuracy compromisation	Complexity improvement of real-time learning with diverse classifiers
Zhang et al. [59]	NSL-KDD, UNSW-NB15	IDS for resource-constrained clustered WSNs	A hierarchical WSN IDS model MK-ELM.	high detection rate, less detection time	Different kinds of attacks detection in WSN considering overall performance improvement
Moustafa et al. [60]	UNSW-NB15	Attack detection in IoT networks	An ensemble of DT, NB, and ANN with the AdaBoost technique to classify features suggested from the DNS, MQTT, and HTTP protocols.	Effective processing of data sources with slightly different feature vectors via error computation.	known as well as unknown attacks detection with significant features from other IoT protocols
Alkadi et al. [61]	UNSW-NB15, Bot-IoT	Privacy preservation and cyber-attacks identification in cloud and IoT background	DBF for data privacy and cyber-attacks identification with Blockchain and BiLSTM techniques.	simple, transparent and safe data exchange. decision support tool for secure data migration	Evaluation of the utility and scalability of DBF with real-world datasets.
Koroniotis et al. [62]	Bot-IoT, UNSW NB15	tracing back the origin of cyber-attack throughout the obfuscated and encrypted network traffic.	Network data flow extraction and integrity verification for encrypted network treatment, Automatic adaption of deep learning parameters using PSO algorithm, development of PSO based DNN to detect and trace anomalous IoT network events.	high detection accuracy of anomalous IoT network events	Parallel computing and GPU for the particles training. Real-world IoT network deployment for different smart systems.
Choudhary et al. [63]	KDD-Cup'99, NSL-KDD, UNSW-NB15	Fast and accurate malicious activity detection in IoT network	DNN for attack detection in IoT network	Higher accuracy rate with all concerned datasets	Use of different DNN architectures

Figure 5. NIDS in advanced networks.

context [61]. A deep learning-based CIDS model is built to cope with security threats imposed on the cloud and the IoT systems from migrated network data. This model employs the Bidirectional Long Short-Term Memory (BiLSTM) technique to develop the CIDS. The performance of the model is evaluated using UNSW-BN15 and BoT-IoT datasets. Smart contracts and Blockchain techniques offer protection to distributed intrusion detection engines. This framework offers a decision support tool for secure data migration. Simple, transparent, and safe data exchange are the outcomes of this experiment.

Another key IoT network security challenge is tracing back the origin of cyber-attack throughout the obfuscated and encrypted network traffic. A novel Particle Deep Framework (PDF) recognizes and traces attack patterns in different phases of digital investigation [62]. This PDF comprises three tasks: (1) Network data flow extraction and integrity verification for encrypted network treatment; (2) Automatic adaption of deep learning parameters using Particle Swarm Optimization (PSO) algorithm; and (3) development of PSO based DNN to detect and trace anomalous smart homes IoT network events. This framework achieves high detection accuracy of anomalous IoT network events. This framework can be further improved by utilizing parallel computing and GPU for the training of the particle to reduce training time.

Fast and accurate malicious activity detection in the IoT network is another challenge. DNN is a preferred method for attack detection in such a scenario [63]. This technique achieves a higher accuracy rate with all concerned datasets, namely KDD-Cup'99, NSL-KDD, and UNSW-NB15. Different DNN architectures will be considered in the future for the proposed method.

Figure 5 summarizes the above mentioned NIDS models.

5. Problems associated with NIDS

5.1. The higher false detection rate of IDS

Anomaly-Based NIDS wrongly categorizes normal but previously unseen system activities as an anomaly. This increases the FP rate. On the other hand, the reason behind the increase in the false-negative rate is the high frequency of new attacks introduced in cyberspace nowadays. Signature-based NIDS stores known attack signatures, and cannot detect new attacks [64,65]. Moreover, some signature-based NIDS may be so specific that a mild change in attack can avoid its detection. In such situations, security experts have no awareness that an attack took place. False negatives cannot be easily judged. Theoretically, a blend of signature-based detection and anomaly-based detection approaches is supposed to be an improvement over both the single approaches. But in the case of a hybrid approach where an anomaly detector creates a list of anomalous observations that are further classified by a signature-based detector into known attacks. In such a case, if the anomaly detector fails to detect an attack because of its similarity with normal behavior patterns, it cannot be detected by the signature-based detector in a later stage [66].

5.2. IDS evaluation with large real-time network traffic

Drastically increased internet data and users making it a puzzling task to monitor huge real-time network traffic. It is essential for the improvement of IDSs to thoroughly analyze both normal as well as abnormal traffic behavior. Learning and detecting attack patterns accurately from such huge data requires a huge amount of training data for generating better results. Modeling and evaluating IDSs with large real network traffic is one of the current key challenges.

5.3. Inefficient intrusion datasets

Enormous unknown patterns of network intrusions are detected recently which are still growing in count continuously at a rapid rate. Therefore updating intrusion datasets periodically is a necessity. This will help in representing appropriate architecture for testing old as well as recently observed network anomalies. A big concern in a multi-cloud environment is the unavailability of datasets for recent security attack analysis, due to privacy issues [67]. The unavailability of efficient intrusion datasets comprising an adequate amount of relevant intrusion types is a big issue.

5.4. Data imbalance

Uneven record distribution in imbalanced Datasets leads to the biased classification of records. The detection rate of a class with fewer records is very less as compared to the detection rate of a class with a majority of records [68]. A variety of data balancing techniques are available to convey this issue but at the cost of increased computational complexity and execution time complexity.

5.5. Slow learner IDS

Slow learner IDS is another issue usually ignored. But it should be dealt with with proper attention to fulfill the need for current requirements of the big data situation [69]. Timely detection of intrusion can save target systems/organizations from massive damage.

5.6. Tedious class labeling in supervised learning IDS approaches

Class labeling is very tedious for field experts when a dataset reaches multi-gigabytes or even more in size

5.7. IDS protection

Besides IDS by itself is prone to be attacked [70]. ML-based IDS models learn attack patterns from the input data. Such models can be a victim of the adversarial attack, wherein minor modification can disguise the traffic classifier [71]. Watermarking techniques are gaining popularity in protecting software against cyber-attacks [72]. Such techniques can be utilized for IDS protection.

6. Conclusion and future scope

This paper focuses on important aspects in the area of network intrusion detection. Many NIDS are built using publically available intrusion datasets. This paper highlights the characteristics and limitations of a variety of publicly available intrusion datasets including the Botnet dataset and Malware datasets. This has resulted in a better understanding of the nature and area of applications of these datasets. It also concludes that there is a need to update intrusion datasets and generate new comprehensive, and efficient datasets. Important aspects of ML techniques are discussed with their application on intrusion detection. ML techniques are competent to handle a large amount of evolving and complex data, but, these techniques have their characteristics and limitations that are to be considered before building a NIDS model. This paper also presents a study of recent NIDS models that exploited the ML-techniques and public intrusion datasets. Different networking environments are considered to conduct this survey. This study presents a clear vision of the current security challenges, solutions, outcomes, and future directions. Hence, this work will be helpful to the researchers to identify a suitable dataset and ML techniques for effective IDS modeling in different networking environments for carrying out their research. Further, this paper can be extended to analyze the real-time monitoring of rapidly increasing network traffic which is still a challenge and an interesting topic of current network security researches [73,74].

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Ms. Geeta Singh is an Assistant Professor and Ph.D. candidate in the School of Computer Science and Engineering at Vellore Institute of Technology, Vellore, Tamil Nadu, India. She received her master's degree in Computer Science from Barkatullah University, Bhopal in 2006. She received her master's degree in Software Systems from RGPV, Bhopal in 2014. Her research interest covers Information Security and Machine Learning Techniques.

Dr. Neelu Khare is presently working as Associate Professor in the School of Information Technology and Engineering at VIT University, Vellore, Tamil Nadu, India. She completed her Ph.D. degree from MANIT Bhopal, India. She has published 41 papers in International Journals and conferences. She guided 4 Ph.D. students. Her areas of interest are Data Mining: Association, Classification, Soft computing techniques, Security, Machine Learning, IoT, and Bio-informatics.

ORCID

Geeta Singh http://orcid.org/0000-0001-7540-5537
Neelu Khare http://orcid.org/0000-0001-9516-0637

References

[1] Di Pietro R, Mancini LV. *Intrusion detection systems* (Vol. 38). New York: Springer Science & Business Media; 2008.

[2] Gu G, Fogla P, Dagon D, et al. Measuring intrusion detection capability: an information-theoretic approach. In: Ferngching Lin, editor. Proceedings of the 2006 ACM symposium on information, computer and communications security; New York: Association for Computing Machinery; 2006 Mar. p. 90–101.

[3] Axelsson S. The base-rate fallacy and the difficulty of intrusion detection. ACM Trans Inf Syst Secur (TISSEC). 2000;3(3):186–205.

[4] Gartner Survey. Newsroom. 2019. [cited 2019 Jan 23]. Available from: https://www.gartner.com/en/newsroom/press-releases/2019-01-23-gartner-survey-finds-government-cios-to-focus-technol.

[5] Hoque N, Bhuyan MH, Baishya RC, et al. Network attacks: taxonomy, tools and systems. J Netw Comput Appl. 2014;40:307–324.

[6] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutorials. 2015;18(2):1153–1176.

[7] Ahmed M, Mahmood AN, Hu J. A survey of network anomaly detection techniques. J Netw Comput Appl. 2016;60:19–31.

[8] Ring M, Wunderlich S, Scheuring D, et al. A survey of network-based intrusion detection data sets. Comput Secur. 2019;86:147–167.

[9] Divekar A, Parekh M, Savla V, et al. Benchmarking datasets for anomaly-based network intrusion detection: KDD CUP 99 alternatives. 2018 *IEEE 3rd International Conference on Computing,* Communication and Security (ICCCS); 2018 October; IEEE. p. 1–8.

[10] Lippmann R, Haines JW, Fried DJ, et al. The 1999 DARPA off-line intrusion detection evaluation. Comput Netw. 2000;34(4):579–595.

[11] Elkan C. Results of the KDD'99 classifier learning contest. Sponsored by the International Conference on Knowledge Discovery in Databases; 1999 September.

[12] Engen V, Vincent J, Phalp K. Exploring discrepancies in findings obtained with the KDD Cup'99 data set. Intell Data Anal. 2011;15(2):251–276.

[13] Tavallaee M, Bagheri E, Lu W, et al. A detailed analysis of the KDD CUP 99 data set. 2009 IEEE symposium on computational intelligence for security and defense applications; 2009 July; IEEE. p. 1–6.

[14] McHugh J. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by Lincoln laboratory. ACM Trans Inf Syst Secur (TISSEC). 2000;3(4):262–294.

[15] Devan P, Khare N. An efficient XGBoost–DNN-based classification model for network intrusion detection system. Neural Comput Appl. 2020;32:1–16.

[16] Khare N, Devan P, Chowdhary CL, et al. SMO-DNN: spider monkey optimization and deep neural network hybrid classifier model for intrusion detection. Electronics (Basel). 2020;9(4). https://doi.org/10.3390/electronics9040692.

[17] Kolias C, Kambourakis G, Stavrou A, et al. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset. IEEE Commun Surv Tutorials. 2015;18(1):184–208.

[18] Laptev N, Amizadeh S. Yahoo anomaly detection dataset s5. 2015. Available from: http://webscope. *sandbox. yahoo. com/catalog. php.*

[19] Lavin A, Ahmad S. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. 2015 *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*; 2015 December; IEEE. p. 38–44.

[20] Singh N, Olinsky C. Demystifying Numenta anomaly benchmark. *2017 International Joint Conference on Neural Networks (IJCNN)*; 2017 May; IEEE. p. 1570–1577.

[21] Song J, Takakura H, Okabe Y. Description of kyoto university benchmark data. 2006. [cited 2016 Mar 15]. Available from: http://www.takakura.com/Kyoto_data/BenchmarkData-Description-v5. pdf.

[22] Song J, Takakura H, Okabe Y, et al. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*; 2011 April. p. 29–36.

[23] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *Military communications and information systems conference (MilCIS)*, 2015. New York: IEEE; 2015. p. 1–6.

[24] Moustafa N, Slay J. The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. Inf Secur J: A Global Perspect. 2016;25(1-3):18–31.

[25] Koroniotis N, Moustafa N, Sitnikova E, et al. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. Future Gener Comput Syst. 2019;100:779–796.

[26] Taheri R, Ghahramani M, Javidan R, et al. Similarity-based Android malware detection using hamming distance of static binary features. Future Gener Comput Syst. 2020;105:230–247.

[27] Zhou Y, Jiang X. Dissecting android malware: characterization and evolution. 2012 *IEEE symposium on security and privacy*; 2012, May; IEEE. p. 95–109.

[28] Arp D, Spreitzenbarth M, Hubner M, et al. Drebin: effective and explainable detection of android malware in your pocket. In: *Ndss*. Vol. 14.. San Diego: Internet Society; 2014 Feb. p. 23–26.

[29] Contagio Dataset. 2020. [ctied 2020 Oct 21] Available from: http://contagio minidump.blogspot.com/. https://www.sec.cs.tu-bs.de/ ∼ danarp/drebin/.

[30] Ghahramani Z. Unsupervised learning. In: Bousquet O, Von Luxburg, Rätsch G, editors. *Summer school on machine learning*. Berlin, Heidelberg: Springer; 2003 Feb. p. 72–112.

[31] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–674.

[32] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[33] Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv (CSUR). 1995;27(3):326–327.

[34] Deng L, Yu D. Deep learning: methods and applications. Found Trends® Signal Process. 2014;7(3–4):197–387.

[35] Yassin W, Udzir NI, Muda Z, et al. A cloud-based intrusion detection service framework. *Proceedings Title: 2012 International Conference on Cyber Security*, Cyber Warfare and Digital Forensic (CyberSec); 2012 June; IEEE. p. 213–218.

[36] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–999.

[37] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: Boulicaut JF, Esposito F, Giannotti F, et al., editors. *European conference on machine learning*. Berlin, Heidelberg: Springer; 2004 Sept. p. 39–50.

[38] Hawkins J, Ahmad S, Dubinsky D. Hierarchical temporal memory including HTM cortical learning algorithms. *Techical report*. Palto Alto: Numenta, Inc; 2010. Available from: http://www. numenta. com/htmoverview/education/ HTM_CorticalLearningAlgorithms. pdf.

[39] Twitter ADVec. [cited 2017 Jan 20]. Available from: https://github.com/twitter/AnomalyDetection.

[40] Stanway A. etsy/skyline [Online code repository]. 2013. Available from: https://github.com/etsy/skyline.

[41] Thill M, Konen W, Bäck T. Online anomaly detection on the webscope S5 dataset: A comparative study. In: Igor Skrjanc, Saso Blazic, editors. *Evolving and adaptive intelligent systems (EAIS)*, 2017. Ljubljana: IEEE; 2017 May. p. 1–8.

[42] Lopez-Martin M, Carro B, Sanchez-Esguevillas A. Application of deep reinforcement learning to intrusion detection for supervised problems. Expert Syst Appl. 2020;141:112963.

[43] Mishra S, Sagban R, Yakoob A, et al. Swarm intelligence in anomaly detection systems: an overview. Int J Comput Appl. 2018: 1–10.

[44] Reddy GT, Khare N. FFBAT-optimized rule based fuzzy logic classifier for diabetes. Int J Eng Res Afr. 2016;24:137–152. Trans Tech Publications Ltd.

[45] Reddy GT, Khare N. Hybrid firefly-bat optimized fuzzy artificial neural network based classifier for diabetes diagnosis. Int J Intell Eng Syst. 2017a;10(4):18–27.

[46] Reddy GT, Khare N. Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction. Int J Fuzzy Syst Appl (IJFSA). 2017b;6(2):25–42.

[47] Reddy GT, Khare N. An efficient system for heart disease prediction using hybrid OFBAT with rule-based fuzzy logic model. J Circuits Syst Comput. 2017c;26(04):1750061.

[48] Tama BA, Comuzzi M, Rhee KH. TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. IEEE Access. 2019;7:94497–94507.

[49] Hamid Y, Sugumaran M. A t-SNE based non linear dimension reduction for network intrusion detection. Int J Inf Technol. 2020;12(1):125–134.

[50] Khan FA, Gumaei A, Derhab A, et al. A novel two-stage deep learning model for efficient network intrusion detection. IEEE Access. 2019;7:30373–30385.

[51] Yang Y, Zheng K, Wu C, et al. Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network. Sensors. 2019;19(11):2528.

[52] Iwendi C, Khan S, Anajemba JH, et al. The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems. Sensors. 2020;20(9):2559.

[53] Vasilomanolakis E, Karuppayah S, Mühlhäuser M, et al. Taxonomy and survey of collaborative intrusion detection. ACM Comput Surv (CSUR). 2015;47(4):1–33.

[54] Djenouri D, Khelladi L, Badache AN. A survey of security issues in mobile ad hoc and sensor networks. In: Dusit Niyato, editor. *IEEE Communications surveys Tutorials*. Vol. 7, No. 4Singapore: IEEE Communications Society; 2005. p. 2–28.

[55] Pathan ASK, ed. Security of self-organizing networks: MANET, WSN, WMN, VANET. CRC press; 2016.

[56] Zhang T, Zhu Q. Distributed privacy-preserving collaborative intrusion detection systems for VANETs. IEEE Trans Signal Inf Process Over Netws. 2018;4(1):148–161.

[57] Gao Y, Wu H, Song B, et al. A distributed network intrusion detection system for distributed Denial of service attacks in vehicular Ad Hoc network. IEEE Access. 2019;7:154560–154571.

[58] Hassine K, Erbad A, Hamila R. Important complexity reduction of random forest in multi-classification problem. 2019 *15th International Wireless Communications & Mobile Computing Conference (IWCMC)*; 2019 June; IEEE. p. 226–231.

[59] Zhang W, Han D, Li KC, et al. Wireless sensor network intrusion detection system based on MK-ELM. Soft comput. 2020;24:1–14.

[60] Moustafa N, Turnbull B, Choo KKR. An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things. IEEE Internet Things J. 2018;6(3):4815–4830.

[61] Alkadi O, Moustafa N, Turnbull B, et al. A deep blockchain framework-enabled collaborative intrusion detection For protecting IoT and cloud networks. IEEE Internet Things J. 2020.

[62] Koroniotis N, Moustafa N, Sitnikova E. A new network forensic framework based on deep learning for internet of Things networks: A particle deep framework. Future Gener Comput Syst. 2020;110:91–106.

[63] Choudhary S, Kesswani N. Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT. Procedia Comput Sci. 2020;167:1561–1573.

[64] Hajisalem V, Babaie S. A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection. Comput Netw. 2018;136:37–50.

[65] Lee W, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat*. No. 99CB36344); 1999, May; IEEE. p. 120–132.

[66] Tombini E, Debar H, Mé L, et al. A serial combination of anomaly and misuse IDSes applied to HTTP traffic. *20th Annual Computer Security Applications Conference*; 2004 December; IEEE. p. 428–437.

[67] Salman T, Bhamare D, Erbad A, et al. Machine learning for anomaly detection and categorization in multi-cloud environments. 2017 *IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*; 2017, June. IEEE. p. 97–103.

[68] Patel H, Singh Rajput D, Thippa Reddy G, et al. A review on classification of imbalanced data for wireless sensor networks. Int J Distrib Sens Netw. 2020;16(4):155014772091640.

[69] Tan Z, Nagar UT, He X, et al. Enhancing big data security with collaborative intrusion detection. IEEE Cloud Computing. 2014;1(3):27–33.

[70] Wang Z. Deep learning-based intrusion detection with adversaries. IEEE Access. 2018;6:38367–38384.

[71] Taheri R, Javidan R, Pooranian Z. Adversarial android malware detection for mobile multimedia applications in IoT environments. Multimed Tools Appl. 2020: 1–17.

[72] Iwendi C, Jalil Z, Javed AR, et al. Keysplitwatermark: zero watermarking algorithm for software protection against cyber-attacks. IEEE Access. 2020;8:72650–72660.

[73] Bär A, Finamore A, Casas P, et al. Large-scale network traffic monitoring with DBStream, a system for rolling big data analysis. 2014 *IEEE International Conference on Big Data (Big Data)*; 2014, October; IEEE. p. 165–170.

[74] Habeeb RAA, Nasaruddin F, Gani A, et al. Real-time big data processing for anomaly detection: a survey. Int J Inf Manage. 2019;45:289–307.