Computer Communications 153 (2020) 406–440

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Machine Learning Models for Secure Data Analytics: A taxonomy and threat model

Rajesh Gupta [a], Sudeep Tanwar [a], Sudhanshu Tyagi [b], Neeraj Kumar [c,d,e,*]

[a] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India
[b] Department of Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Deemed to be University, Patiala, Punjab, India
[c] Department of Computer Science Engineering, Thapar Institute of Engineering and Technology, Deemed to be University, Patiala, Punjab, India
[d] Department of Computer Science and Information Engineering, Asia University, Taiwan
[e] King Abdul Aziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Keywords:
Big data
Secure Data Analytics
Data reduction
Machine learning models
Threat model
Data security and privacy

ABSTRACT

In recent years, rapid technological advancements in smart devices and their usage in a wide range of applications exponentially increases the data generated from these devices. So, the traditional data analytics techniques may not be able to handle this extreme volume of data known as Big Data (BD) generated by different devices. However, this exponential increase of data opens the doors for the different type of attackers to launch various attacks by exploiting various vulnerabilities (SQL injection, OS fingerprinting, malicious code execution, etc.) during data analytics. Motivated from the aforementioned discussion, in this paper, we explored Machine Learning (ML) and Deep Learning (DL)-based models and techniques which are capable off to identify and mitigate both the known as well as unknown attacks. ML and DL-based techniques have the capabilities to learn from the traffic pattern using training and testing datasets in the extensive network domains to make intelligent decisions concerning attack identification and mitigation. We also proposed a DL and ML-based Secure Data Analytics (SDA) architecture to classify normal or attack input data. A detailed taxonomy of SDA is abstracted into a threat model. This threat model addresses various research challenges in SDA using multiple parameters such as-efficiency, latency, accuracy, reliability, and attacks launched by the attackers. Finally, a comparison of existing SDA proposals with respect to various parameters is presented, which allows the end users to select one of the SDA proposals in comparison to its merits over the others.

Contents

* Corresponding author.
E-mail addresses: 18ftvphde31@nirmauni.ac.in (R. Gupta), sudeep.tanwar@nirmauni.ac.in (S. Tanwar), s.tyagi@thapar.edu (S. Tyagi), neeraj.kumar@thapar.edu (N. Kumar).

https://doi.org/10.1016/j.comcom.2020.02.008
Received 29 November 2019; Received in revised form 28 January 2020; Accepted 3 February 2020
Available online 10 February 2020

1. Introduction

With the increasing demand of Quality of Service (QoS) provisions to the end-users using the Internet of Things (IoT) and the Internet of Everything (IoE) technologies, Big Data Analytics (BDA) becomes one of the key areas to be explored by the research community across the globe. The sensors and physical devices, which are referred to as *things* in IoT, are predominately used for sensing, computation, and transmission and are major sources of data generation. However, the usage of these smart devices exponentially increases the data rate generation [1]. It has been observed from the survey that over a while, this data is massively multiplied each second and getting more massive in terms of size. As per the report presented by *EMC* in 2014, the data size will reach up to *44 zettabytes* or *44 trillion gigabytes* till 2020 [2] and in 2025 it will be approximately *163 zettabytes* as per the report published by *Economic Times* [3]. So, the traditional servers are not capable of handling this huge amount of data keeping in view of the data computation and storage constraints.

Also, this enormous amount of data generated from different sources, most of which are connected to the Internet, open the doors for the attackers to launch various types of attacks (for example, cross-site scripting, ransomware, SQL injections, etc.). The traditional security mitigation techniques may not be applicable to detect these threats because of low accuracy with respect to known and unknown attacks [4]. However, Machine Learning (ML)-based models and techniques can be a viable solution to mitigate the aforementioned security threats. ML algorithms can process large datasets to extract meaningful information from the database repository. Moreover, ML algorithms can update the models in real-time as per their requirements and are used to mine and process BD intelligently and efficiently. But sometimes, the security of the systems may be compromised due to massive data processing and analysis.

Most of the organizations are spending or willing to pay millions of dollars to secure their infrastructure, but the attackers can attack their confidential data despite traditional secured infrastructure. In such a scenario, ML models are more efficient and intelligent to identify any abnormality as compared to humans, especially in the era of rapidly growing data generation rate. These models have been trained on some pre-determined guidelines/rules to similar patterns, relationships between various parameters such as type of the organization, user behavior, and transaction. Through this training, it is possible to identify any abnormal traffic for the network. Moreover, these ML models help the organizations to early detect the attacks and breaches and give enough time to the organizations to identify and mitigate various threats.

In ML models, features are to be extracted manually by humans and feed it into the classification algorithm for data classification. There may be a chance that humans may miss or skip some or few features, which result in improper data classification. In 2006, Professor Hinton proposed the Deep Learning (DL) theory (*subset of ML*) to overcome the aforementioned issues of ML models [5]. In DL, feature extractions are automatic without the intervention of humans. It has the ability to extract a better representation of data by extracting all of its features. It improves the accuracy and believability of systems in detecting attacks and intruders.

ML and BD have already set the stage for the researchers to redefine the security schemes as per the latest trend in the technology. ML algorithm can be used to analyze the traffic patterns and with the help of a decision tree, it can be classified as suspicious or normal data. The suspicious data can be ignored and normal data can be used for further processing. This way, ML algorithms can bifurcate the suspicious data from the incoming data stream and can achieve higher analytics accuracy. But, what to do when the intruders attacked on your ML models during data analytics? And how to detect such security attacks on the designed ML models? Motivated from the aforementioned discussion, in this paper, we explore the applicability of various ML models usage for attack models and mitigation techniques used. Some examples of security attacks during data analytics are as follows:

- *Spam messages filter:* If ML-based models are trained with some false datasets, then it results in the wrong classification of the spam and legitimate messages. It has a direct impact on the system integrity and authenticity of the messages; thus, the system availability is compromised.
- *Bank loan sanction models:* Training of this classification model with some compromised datasets results in the classification of the loan defaulters as good users. Moreover, its execution results in business loss to banks due to the approval of a loan to the defaulters.
- *Financial credit risk model:* Training with false datasets results in the wrong classification of the business/buyers. So, there are high credit risk exposures in such an environment.
- *Life Insurance model:* Inappropriately trained life insurance model may lead to offer higher discounts to a group of users resulting in the approval of the insurance to those who are not even entitled or qualified for it.

1.1. Scope of this survey

Many of the surveys explored by different authors on the security of data analytics with ML algorithms are as follows [21,6–16]. Most of these surveys, as per our knowledge, have focused on data security and input network traffic analysis with data mining, ML, and DL techniques. But, no exhaustive survey exists which explored the SDA with ML and DL algorithms to provide threats identification and the mitigation techniques used. The proposed survey covers supervised, unsupervised, ensemble, semi-supervised, reinforcement, and DL techniques for SDA. One of the surveys by Singh et al. [21] highlighted the security concerns of BDA with ML techniques to understand network behavior. But, they have not discussed the mitigation strategies. Similarly Mayhew et al. [6] explored the suitability of ML algorithms in BDA to mitigate the insider attack.

Choudhary et al. [8] discussed the ML-based network security solution for BDA. Later on, Yavanoglu et al. [11] tested the above ML-based solution over the dataset, which contains both suspicious and normal data. They found it effective against network traffic and IDS. Habeeb et al. [15] explored the network anomaly detection based on ML algorithm in BDA. But, they have not focused on computation cost and redundancy of data which may affect the performance of the proposed system.

Gardiner et al. [9] presented the comparison of various state-of-the-art ML clustering and classification algorithms for possible attacks detection, but no framework or architecture was proposed for attack mitigation. Singh et al. [10] presented a survey on ML algorithm challenges in processing BD over both labeled and unlabeled datasets to resolve privacy issues. Mishra et al. [13] presented a detailed survey on ML models to detect multiple low-frequency attacks. Hu et al. [7] evaluated the energy associated with BDA and associated security and privacy issues. Their approach managed to detect spoofing, man-in-the-middle, and Denial-of-Service (DoS) attacks.

Jiang et al. [12] proposed an LSTM-RNN based intelligent attack detection scheme to detect data that is unusual or not. This DL-based approach increases the accuracy of attack detection. Husak et al. [14] surveyed both discrete and continuous ML models to detect network intruders. Mahdavinejad et al. [16] presented the ML algorithm applicability in smart city for data analytics. Table 1 presents the summary of these surveys and their differences in comparison to the proposed survey.

1.2. Motivation

The main motivation of this paper is as follows:

- The existing literature is mainly focused on data security at the data processing or collection phases using traditional cryptographic techniques. These techniques are more computing extensive and not able to defend against the attacks which are not in their database. So, there is a need for a survey that discusses the prediction modeling for future attacks which are not in its database repository and accordingly generate the defense mechanism.
- No survey has discussed the taxonomy for security attacks on data analytics as well as ML and DL-based mitigation solutions in a single paper.
- Using this review, researchers will get motivated and get an insight to the secure data analytics using ML and DL models.

1.3. Contribution of this survey

In this paper, we present an in-depth review on SDA using various ML and DL models. We also highlight the various challenges in SDA implementation and their countermeasures. Then, we explored "*How ML and DL models can be used for intrusion detection and mitigation?*". Based on the above discussion, the following are the contributions of this paper.

- We present a comprehensive and systematic review of SDA using ML and DL models by exploring the possible security threats on data analytics.
- Then, we proposed an SDA-ML architecture based on both supervised and unsupervised learning algorithms.
- Finally, we provide an in-depth survey of the research challenges for SDA implementation in context with BD exploration by different devices.

1.4. Organization and reading map

The structure of the survey is as shown in Fig. 1. Table 2 lists all the acronyms used in this paper. Section 2 provides the background of SDA. In Section 3, we highlighted the survey methodology used for the survey. In Section 4, we discussed the security attacks in data analytics. In Section 5, we discussed the SDA architecture and threat model. In Section 6, we discussed the case study of smart grid in data analytics. In Section 7, taxonomy of ML and DL based SDA. In Section 8, we discussed the open issues and research challenges in SDA implementation, and finally, Section 9 concludes the paper.

Fig. 2 provides the reading map. Readers having interests in the basics of SDA can focus on Sections 1, 2, and 6. The survey methodology is given in Section 3. The architecture and threat model of SDA are given in Sections 1 and 4. Readers with interest in the usage of ML and DL in SDA can focus their reading on Sections 1 and 5. Finally, we recommend Sections 2, 3, 4, 5, and 6 to the readers interested in gaining a high-level overview of SDA including open issues and research challenges.

This section gives the glimpses of possible attacks on data generated from the different sources and their mitigation techniques using ML and DL models. It also presents the comparison of existing surveys present in SDA using different parameters. In this section, readers can easily find the original contributions of this paper and the organization of the article along with the reading map.

Table 1
Comparison between the proposed and existing secure data analytics surveys.

Year	Authors	Objectives	Major Contributions	Related contents in proposed survey	Attacks detected	Dataset used to test the approach	Merits	Demerits
2015	Mayhew et al. [6]	To explore the suitability of ML in BDA in presence of the insider attacks	–	4.3.1 and 4.3.2	Zero day attack	HTTP, TCP, WiKi, Twitter, Email	Behavior-Based Access Control model used for attack detection	Only few attacks can be detected using mentioned technique
2016	Hu et al. [7]	To explore the big energy data analytics and its associated security issues	-Presented a survey on energy big data analytics and security - Describe the taxonomy of big data analytics and security issues	–	Profiling, spoofing, man-in-the-middle, DoS, and system hijacking attack	Smart-grid dataset	Exhaustive taxonomies are given to highlight the relationships among different variables during SDA	Most of the key parameters used during SDA are missing
2016	Chaudhari et al. [8]	To explore the ML based solutions for security issues and associated challenges in BD	- Reviewed the security concerns of big data	4.3	Anomaly detection	–	Presented the various data mining privacy preserving techniques for SDA	Not provided solutions to mitigate different attacks
2016	Gardiner et al. [9]	Presented the comparison of state-of-the-art ML algorithms with possible attacks	- Review on C&C detection - Identify the weakness of the C&C system	4.3.1 and 4.3.2	Bridging attack, gradient descent attack	MNIST handwritten digits dataset for attack demonstration	Discussed the possible attacks on classification and clustering algorithms	No framework for attacks modeling is given to overcome security issues
2017	Singh et al. [10]	Presented a comprehensive survey on challenges faced by ML algorithms in processing the BD	- Discusses the issues related to mining big data in ML perspective	5	–	Both labeled and unlabeled datasets	Focused on streaming high-dimensional data	Focused on privacy issues only not other attack issues
2017	Yavanoglu et al. [11]	They focused on AI and ML algorithms datasets to analyze and detect network traffic and abnormalities in BDA	- A comprehensive review on publicly available datasets - Assessment of AI and ML models using datasets	4.3.1 and 4.3.2	Cybersecurity, network traffic, and IDS	KDD, ISOT, CSIC, CTU-13, ADFA, and UNSW-NB15	Highlighted the pros and cons of ML datasets (KDD Cup and CTU-13) in view of network traffic	Solutions to mitigate the attacks are not discussed.
2018	Jiang et al. [12]	Proposed a Long Short Term Memory (LSTM)-Recurrent neural network (RNN) based intelligent attack detection architecture, i.e., whether the input is an attack or normal	- DL-based attack detection in social networks methodology - Designed voting algorithm to identify the input stream is attack or not	4.3.1	Network traffic	NSL-KDD	More accuracy in comparison to other ML algorithms with respect to detection rate, accuracy and false alarm rate	It mainly focused on ML-based network traffic attack detection but not explored other types of attacks
2018	Mishra et al. [13]	Presented a detailed survey on capabilities of ML techniques to detect various low-frequency security attacks with network attack dataset	- Classified the attacks as per the characteristics - Discussion on intrusion detection based on existing literature	4.3, 4.3.1 and 4.3.2	Fuzzers, DoS, worms, analysis, generic, shell code and back-door attacks	KDD-99 DARPA dataset	Optimal feature selection to detect multiple attacks	More ML-based security solutions need to be explored and the performance of ML techniques is not considered
2018	Husak et al. [14]	To have a comparative analysis on discrete and continuous models to detect intrusive activities	- What can be predicted in a cyber security domain? - The usability of predictions in cyber security?	–	Malicious, malware, back-door, enumeration attacks	DARPA grand challenge problem dataset	Discussed the applicability of ML methods and related problems	Not applicable for collaborative intrusion detection environment
2018	Habeeb et al. [15]	To explore the ML algorithms for real-time BD processing to detect anomalies	- Surveyed the literature focused on big data processing for anomaly detection	4.3, 4.3.1 and 4.3.2	Anomaly detection	KDD, HTTP proxy log dataset	Detects anomalies in BD with high accuracy	Did not explore the computation cost, redundancy, and parameter selection

(continued on next page)

Table 1 (*continued*).

Year	Authors	Objectives	Major Contributions	Related contents in proposed survey	Attacks detected	Dataset used to test the approach	Merits	Demerits
2018	Mahdavinejad et al. [16]	Presented a taxonomy of ML algorithms to be used in smart city to extract insights from the collected data	- How ML algorithms can be used in IoT smart data? - Taxonomy of ML models adopted in IoT	4	–	Intel Lab dataset	Highlighted the key challenges of ML for IoT data analytics	Attacks models are not explored
2019	Liu et al. [17]	Presented a good survey on SDA but have focused only on Edge computing	- Analyzed the security threats on data analytics - Proposed the security and performance requirements - Highlighted open issues and future directions	4.1, 4.2, 4.3	Jamming, DoS, Privacy Leakage, Data Injection attacks, Physical attacks	–	Discussed all possible attacks on data analytics	No framework proposed to achieve SDA.
2019	Moghadam et al. [18]	Highlighted the cryptographic and security mechanisms to ensure SDA in cloud environment	- Compare practical secure solutions presented in literature - Given a system model that uses cloud data analytics	4.1, 4.2, 4.3	linking, inference, and frequency analysis attacks	Big datasets	Described the encrypted-based solutions for data analytics	Detects the attacks which are in its database definition and cannot predict the future attacks
2019	Kumari et al. [19]	Analyzed the streaming of big data generated from varied IoT devices and highlighted the challenges in processing heterogeneous data	- Architecture for securing streaming big data - Analysis of existing surveys on big data streaming	4.1, 4.2, 4.3	DoS, malware, and phishing attacks	Big data environment	Presented a taxonomy and the case study of securing big data analytics in healthcare 4.0	Not discussed the types of attacks on data analytics in details
2019	Pitropakis et al. [20]	Presented a taxonomy of possible attacks against the ML techniques	- Taxonomy of adversarial ML (AML) approaches - Review on AML towards categorization and classification	4.1, 4.2, 4.3	–	–	Presented the attacks on ML models and their defenses	Not explained the types of attacks on data analytics during data collection phase and relative comparison on different approaches
2020	Proposed survey	To explore ML models for SDA		–	Known as well as unknown attacks	–	Various attacks on ML models and components of SDA are explored	–

Table 2
A list of abbreviations.

Acronym	Explanation	Acronym	Explanation
AE	Auto Encoders	IoE	Internet of Everything
AES	Advanced Encryption Standard	IoT	Internet of Things
AI	Artificial Intelligence	KNN	K-Nearest Neighbor
ANN	Artificial Neural Network	LSTM	Long Short Term Memory
BD	Big Data	LVQ	Learning Vector Quantization
BN	Bayesian Network	MD	Message Digest
CC	Cloud computing	ML	Machine Learning
CI	Communication Interruption	NeHA	National eHealth Authority
CSI	cybersecurity insurance	NN	Neural Network
DDoS	Distributed DoS	PASTA	Process for Attack Simulation and Threat Analysis
DL	Deep Learning	PCA	Principal Component Analysis
DoS	Denial-of-Service	RF	Random Forest
DPM	Differential Privacy Method	RNN	Recurrent neural network
DT	Decision Tree	SDA	Secure Data Analytics
GM	Gaussian Mixture	SDA-ML	Secure Data Analytics-Machine Learning
GPS	Global Positioning System	SH	Session Hijacking
HEM	Homomorphic Encryption Method	SOM	Self Organization Map
HTTP	HyperText Transfer Protocol	SVM	Support Vector Machine
IDS	Intrusion detection system	VAST	Visual, Agile, and Simple Threat modeling

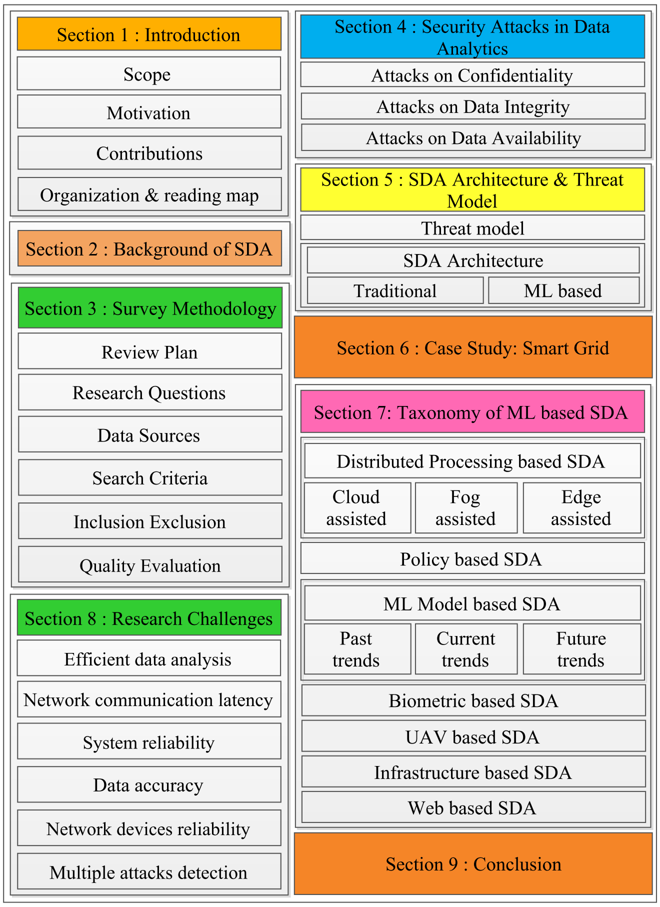

Section 1 : Introduction
Scope
Motivation
Contributions
Organization & reading map
Section 2 : Background of SDA
Section 3 : Survey Methodology
Review Plan
Research Questions
Data Sources
Search Criteria
Inclusion Exclusion
Quality Evaluation
Section 8 : Research Challenges
Efficient data analysis
Network communication latency
System reliability
Data accuracy
Network devices reliability
Multiple attacks detection
Section 4 : Security Attacks in Data Analytics
Attacks on Confidentiality
Attacks on Data Integrity
Attacks on Data Availability
Section 5 : SDA Architecture & Threat Model
Threat model
SDA Architecture
Traditional
ML based
Section 6 : Case Study: Smart Grid
Section 7: Taxonomy of ML based SDA
Distributed Processing based SDA
Cloud assisted
Fog assisted
Edge assisted
Policy based SDA
ML Model based SDA
Past trends
Current trends
Future trends
Biometric based SDA
UAV based SDA
Infrastructure based SDA
Web based SDA
Section 9 : Conclusion

Fig. 1. Structure of the survey article.

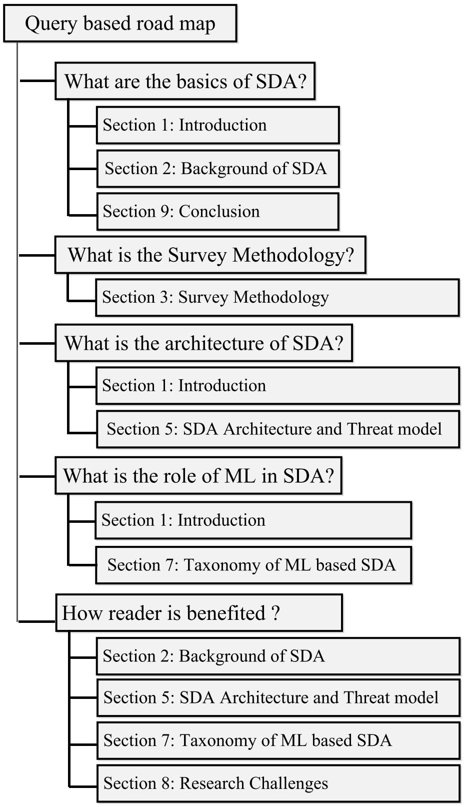

Query based road map
What are the basics of SDA?
Section 1: Introduction
Section 2: Background of SDA
Section 9: Conclusion
What is the Survey Methodology?
Section 3: Survey Methodology
What is the architecture of SDA?
Section 1: Introduction
Section 5: SDA Architecture and Threat model
What is the role of ML in SDA?
Section 1: Introduction
Section 7: Taxonomy of ML based SDA
How reader is benefited ?
Section 2: Background of SDA
Section 5: SDA Architecture and Threat model
Section 7: Taxonomy of ML based SDA
Section 8: Research Challenges

Fig. 2. The reading map.

2. Background of SDA

In the early era of computers, most of the applications were not Internet-enabled. So, security and privacy was not the main issue during data analytics. However, with the advent of the Internet, an open channel, there are possibilities of the launching of various attacks by attackers across the globe. Internet-enabled applications may be vulnerable to the attacks such as-spoofing, snooping, DoS, and SYN flood resulting in compromising of confidentiality, integrity, and availability of the system.

The various attacks on confidentiality, integrity, and availability, along with their categorization are as shown in Fig. 3. Attacks on data availability are divided into two categories (i) Session Hijacking (SH) and (ii) Communication Interruption (CI) attacks. SH attack is a security attack on the user's session using IP spoofing and man-in-the-middle attack. It is categorized into active hijacking and passive hijacking attacks. In the active hijacking attack, the attacker manipulates the client's connection and be-fool the server as the authenticated user, whereas, the attacker in the passive hijacking monitors the network traffic to capture the authenticated information. Various examples of active and passive hijacking attacks are sensor manipulation, spoofing, location manipulation, network traffic monitoring, and sniffing, respectively.

The other category of data availability attack is CI, in which attackers make the network services unavailable to the authenticated users. This attack is against the network service availability. It is classified into three broad categories: (i) DoS attack, (ii) Distributed DoS (DDoS) attack, and (iii) GPS Spoofing. DoS attack makes the network resources inaccessible to the user. It happens with flooding the target. The examples of DoS attacks are ping flood, smurf, buffer overflow, and SYN attack. However, DDoS attackers flooded the communication channel, which restricts the entry of legitimate users. The DDoS attack can be volumetric-based, protocol-based, and application-based. In a volumetric-based attack, an attacker sends the traffic to the target network to saturate its bandwidth capacity. Possible volume-based DDoS attacks are UDP flood, DNS amplification, and NTP amplification attack. In the application-based DDoS attack, attackers target the application layer processes. It is also referred to as layer 7 DDoS attack. Various attacks in this category are HTTP flood, GET/POST flood, and low-and-slow attacks. However, the protocol-based attack makes the target network or node inaccessible by consuming all its processing capabilities. It is an attack on layers 3 and 4 of the protocol stack. Possible protocol-based attacks are SYN flood, ping of death, and smurf attacks.

Attacks on data integrity are divided into two broad categories, such as (i) data fabrication and (ii) data modification. An attacker can either modify the sender data or fabricate the new malicious data to replace the authenticated data. The various data fabrication attacks are SQL injection, route injection, email spoofing, and audit-trail falsification, whereas data modification attacks are replay attack, DNS cache poisoning, website defacement, and malware piggy-backing. The last category of attack is "*attack on data confidentiality*" in which an unauthorized user gets access to the confidential information. The possible attacks on data confidentiality are eavesdropping, man-in-the-middle, and targeted data mining attack.

Threats to the system can be prevented, identified, and mitigated by using threat modeling. It is an iterative process which consists of (i) define enterprise asset, (ii) identify the application with respect to assets, (iii) create security profile, (iv) identify potential security threats, (v) prioritize potential security threats, and (v) document the adverse effects.

This section gives the background knowledge of SDA and various possible attacks during the data analytics. The attacks are explained based on their nature like availability, integrity, and confidentiality. A detailed taxonomy of possible attacks on data analytics is presented in this section.

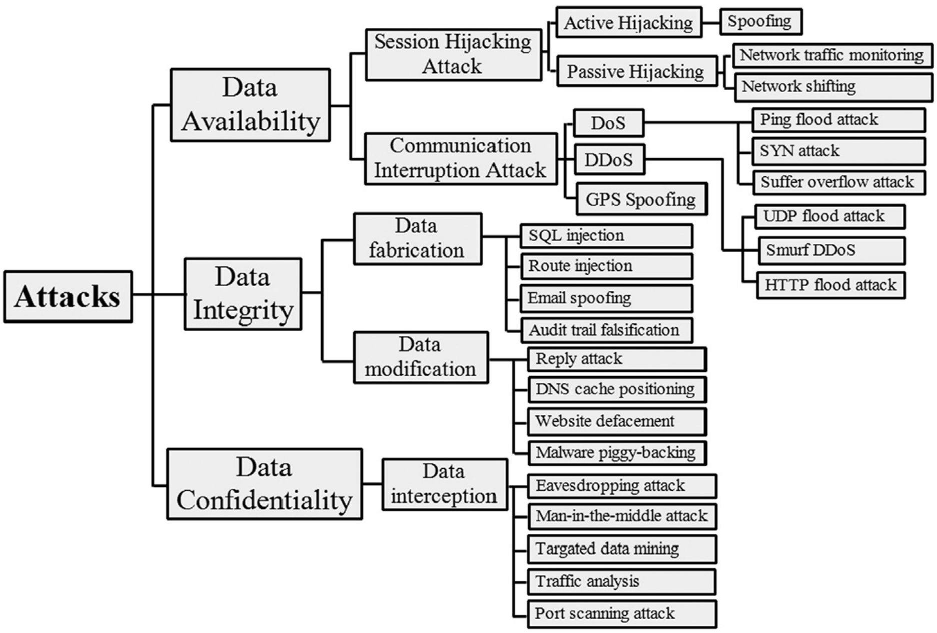


Fig. 3. Categorization of various attacks on the data [22].

3. Survey methodology

This section presents the methodology considered to conduct this survey.

3.1. Review plan

This exhaustive study needs some pre-planning so, we begin with the review planing, identification of numerous research questions (RQ), related data sources, effective search criteria, articles inclusion–exclusion criteria, and quality evaluation. We identified various relevant studies, articles, and publications to carry out this systematic survey. The identified material is first checked for quality before extracting the relevant data for the proposed survey.

3.2. Research questions

In the proposed survey, the authors have identified the relevant literature on security issues in data analytics. Some of the identified RQs along with their objectives used for the systematic survey are listed in Table 3.

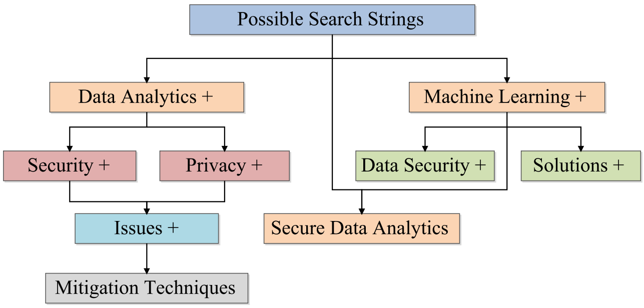


Fig. 4. Search strings.

3.3. Data sources

A broad number of repute and trustable data sources are required for writing a comprehensive survey. We preferred only standard peer-reviewed journal databases such as — IEEEXplore digital library, Science Direct journals (Elsevier), ACM Digital Library, Wiley online library, and Springer link for finding the existing literature on secure data analytics and the electronic data sources recommended by the authors in [23–25].
Other data sources used: Technical books (published in IEEE, Wiley and Springer platforms), reports (Government reports), predicting agencies sites, and online literature (technical blogs) matches with the theme of the proposed comprehensive survey are included.

3.4. Search criteria

In this survey, the search criteria uses the keywords like “Machine Learning for Secure Data Analytics”, (“Security Issues in Data Analytics” AND “Machine Learning in data analytics”) and other related keywords referred for the search criteria in the comprehensive survey as shown in Fig. 4. There exist many research papers/articles in which the search string is not either in title or abstract of the paper, for such cases, a manual search process was executed.

3.5. Criteria of inclusion and exclusion

As the security and privacy are the key challenges in various application areas, so the search string “security and privacy issues” will always give irrelevant papers, that makes the filtration quite tough. To overcome this, the search criteria mentioned in Section 3.4 need to be followed. In order to make this survey attractive and impactful, the recent and relevant papers of 2020 are included along with the early access articles. The other survey articles, tutorial papers, technical patents, books, technical reports, blogs, and other resources are also included for wider coverage. The filtering of papers based on relevancy and other parameters is shown in Fig. 5. This filtration is divided into multiple stages based on the title, abstract, full paper, and investigations. Finally, we identified some relevant papers and focused on those having good citations.

Table 3
Identified research questions and their objectives.

Q. No.	Identified research questions	Objective
RQ.1	What are the Security and privacy issues in data analytics?	It aims to explore the security and privacy issues in data analytics.
RQ.2	Which type of issues exist and how to address those issues?	It aims to search the issues and address those using existing machine learning techniques to predict and ensure the security and privacy in data analytics.
RQ.3	What are the existing studies that are emphasizing on the comparative analysis of existing surveys on secure data analytics or security and privacy issues in data analytics.	It aims to identify the existing surveys in the relevant areas and present a detailed comparison of security and privacy issues in data analytics.
RQ.4	Discuss various research challenges in machine learning-based secure data communication.	It aims to provide information on open issues and research challenges in secure data analytics.
RQ.5	How this study is applicable in a real-life scenario?	It aims to provide the case study of the usability of machine learning techniques in secure data analytics.

Table 4
Quality hiding questions.

Q. No.	Description of question	Answer
Q1	Does the research paper refer to the security and privacy issues in data analytics and its mitigation techniques?	YES
	The papers add an overview of security and privacy issues in data analytics where the word “security and privacy issues” are not being used. Are Such papers excluded from the literature survey?	NO
Q2	Do the abstract, title, and full text of research paper describe the “security and privacy issues of big data analytics”?	YES
	Have the abstract, title, and full text of research paper described the security and privacy issues in data analytics?	NO

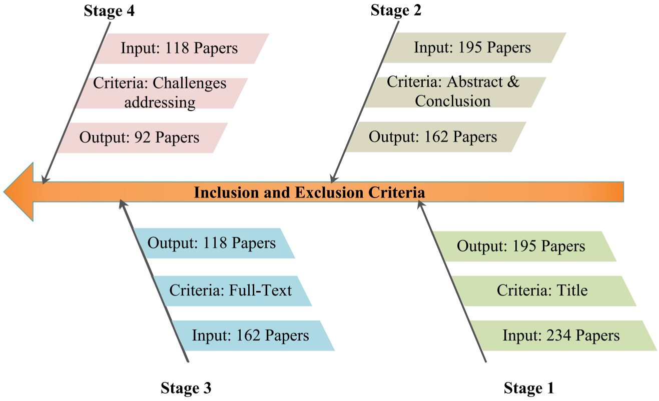


Fig. 5. Criteria of inclusion and exclusion.

3.6. Quality evaluation

In this section, the quality evaluation has been performed on the identified papers as per the guidelines given by the professional bodies, i.e., Database of Abstracts of Reviews of Effects (DARE) and Center for Reviews and Dissemination (CRD) [23]. Quality hiding questions are as given in Table 4 and the quality questions have used to select the research papers.

> This section presents the survey methodology opted for conducting the survey which includes review planing, research questions identified, sources of data from where the literature collected, search criteria used, criteria for article inclusion and exclusion, and the quality checks for the final selection of articles.

4. Security attacks in data analytics

Recent advancements in information and network technologies have created a larger surface for Cybercrimes. It requires faster, efficient, and accurate procedures to detect those cybercrimes. In this section, we discuss various types of possible attacks on data analytics. Then we give a relative comparison of all possible attacks with different parameters.

4.1. Attacks on confidentiality

Data confidentiality means preventing highly confidential information from unauthorized or malicious users. It ensures that only authorized users can access the secured information. Below mentioned are the possible types of confidentiality attacks.

4.1.1. Eavesdropping attack

It is a type of passive attack which is easy to incorporate and hard to detect in the network. In this, an attacker passively listens to the ongoing network communication to access the private information. This may exploit the data processing and analysis information of the server or system. Cryptographic methods are there to prevent communication from the eavesdropping attack, but due to limited power devices such as sensors, it is not suitable.

Defense methods: Ettercap, Kismet, and Wireshark tools are available to defend wireless communication networks against eavesdropping attack [26].

4.1.2. Man-in-the-middle attack

It is like a throwing ball game in which two persons are throwing a ball towards each other, and the third person is trying to catch it. In this, an attacker involved himself in the network to access the messages from the sender, modify it and forward it to the receiver. The third person might have access to the secret or financial analysis information of the sender.

Defense methods: To defend against this attack, we can use secure containers, wrappers, mobile anti-virus, avoid using open Wi-Fi hotspots, and automatic connections.

4.1.3. Traffic analysis attack

In this attack, attackers get the knowledge about data processing and analysis methods and results based upon information eavesdropped by the malicious client. It can be done with dedicated software codes. It then categorizes the network nodes as per the information gathered from them.

Defense methods: Data encryption and channel masking.

4.1.4. Targeted data mining attack

It refers to the threat in which malicious actors actively compromise the target infrastructure with anonymity. It usually conducted in campaigns and attackers can customize, update, and improve their methodology as per the target sector nature and behavior.

Defense methods: Proper & effective data classification, network segmentation, personnel awareness, threat intelligence, and protecting log files.

4.1.5. Port scanning attack

It is a type of attack which finds the active ports among the range of port addresses and exploits the vulnerabilities of the system. It is used to keep track of the data received, data analysis, and data forward activities of the system. The possible types of port scanning are: TCP connect scanning, TCP SYN scanning, TCP FIN scanning, and Fragmentation scanning.

Defense methods: Use of Internet firewalls, filtering ICMP type 3 unreachable data packets, fuzzy rules, and port scan identification tool PortSentry.

4.2. Attacks on data integrity

Data integrity means that the information can be accessed or altered by authorized personals. In this type of attack, an unauthorized person aims to modify the transmitted data intentionally by gaining unauthorized access [27]. They can also target the data analysis results for bringing down the performance of the server or receiver system. Attacks on data integrity are categorized into data fabrication and data modification attacks.

4.2.1. Data fabrication attack

In this type of attack, an unauthorized user intentionally fabricates the illegitimate information (*i.e., counterfeited data*) within a system alongside the authentic information. Attackers use data fabrication to gain the trust of the system as a legitimate user. Various types of data fabrication attacks are:

- *SQL Injection:* Most of the web servers use Structured Query Language (SQL) for managing the data in their databases. So this attack particularly targets those kinds of web servers by executing malicious code to fetch the data from their databases. The data would be, personal customer credentials such as credit card details, id's and passwords.
  Defense methods: Avoid using dynamic SQL, do not keep sensitive data in plain text, limit database privileges & permissions, Sanitizing Inputs, do not display errors to users directly, and use web application firewall.
- *Route Injection:* In this type of attack, an attacker injects additional hops (*i.e., routers*) in the route to the destination with intentions to divert the Internet traffic to the monitored location for analysis and manipulation before it reaches to the destination node. Renesys (an Internet intelligence co.) found approx. 1500 IP addresses hijacked in a year [28].
  Defense methods: Conditional Border Gateway Routing Protocol (BGP) Route Injection.
- *Audit Trail Falsification:* Audit trail is a set of records that provides an evidence of the sequence of all events happened. It generally is financial and healthcare data transactions. In this type of attack, an attacker injects, removes, and manipulates the log entries, which aims to mislead the audit processing.
  Defense methods: Gap in sequence number (record deleted), mismatch hash value (record modified), and duplicate sequence number (record injected)
- *Email Spoofing:* It is the process of forging the source address in the email header. It is used to mislead the email receiver about the source address of the message. It is often used by Spam and worms because email protocols do not have any authentication mechanism.
  Defense methods: Encrypt server email traffic, use email authentication methods (Sender Policy Framework (SPF), Sender ID, Domain Keys Identified Mail (DKIM, and Domain-based Message Authentication, Reporting, and Conformance(DMARC))

4.2.2. Data modification attack

In this attack, some parts of the message either altered or reordered to produce an unauthorized effect. A receiver cannot receive the original message sent by a legitimate user. It may degrade network performance. Various types of data modification attacks are:

- *Replay Attack:* It is a type of playback attack, in which data transmission is either maliciously repeated or delayed. In this, an attacker intercepts the transmitted data and re-transmit it multiple times to affect the data analysis and processing results of the receiver systems. Receivers were unable to get the real data for processing. It also considers as the lower version of a man-in-the-middle attack.
  Defense methods: Use of session ID's, message ID's, use of session tokens, incorporation of message timestamps, and passwords for each transaction or message.
- *DNS Cache Poisoning Attack:* In this attack, a malicious data is put into the cache of Domain Name Server (DNS) to reroute the domain name to the malicious IP address. It may restrict the further processing of data, as the source data may be rerouted to the new IP address. It may also deny the resources or breach the confidentiality of the system.
  Defense methods: Auditing of DNS zones, must install security updates, protect DNS bind version, avoid recursive queries, and use two-factor authentication.
- *Website Defacement Attack:* It is an attack on web applications, which changes the physical or visual appearance of the web page. It can be achieved with the exploitation of a poorly configured system and impacts both website contents and images. Due to this, users may congregate incorrect information that would divert the data processing and analysis results, which leads to false predictions.
  Defense methods: Few tools are available to protect against website defacement attacks are SUCURI, Fluxguard, Status Cake, IPVTec, and WebOrion. Other possible ways are penetration testing, avoid cross-site scripting attacks and security against SQL injection attack.
- *Malware Piggy-backing Attack:* It is a form of wiretapping attack, in which malicious user gains unauthorized access to the system via legitimate user connection if he/she forgot to logout.
  Defense methods: MAC address authentication, ensure IP security (IPSec), and use of a shared key.

4.3. Attacks on data availability

In this type of attack, an attacker aims to make resources and services unavailable for the network nodes or authorized users. The various possible data availability attacks are as shown below.

4.3.1. Session hijacking attack

SH attack is a security attack on the user's session using IP spoofing and man-in-the-middle attack. It comprises of a session token and unauthorized users aim to steal or predict the token to gain unauthorized access. It is categorized into active and passive hijacking attacks.

Active hijacking attack. Attackers in active hijacking attack try to manipulate the client's connection and be-fool the server as an authenticated user. This attack may disrupt the data analysis processing of the system or cloud by injecting a packet or frame and pretends to one of the communication hosts. This is further categorized into:

- *Spoofing Attack:* It was an attack when an attacker proved himself as a legitimate user of the system by falsifying the credential data. The main aim of this attack is to affect organizations reputation, data breach, and revenue loss. It can be of many types: email spoofing, caller ID spoofing, ARP spoofing, and DNS server spoofing.
  Defense methods: Examine the communication legitimacy (spelling, grammar, and tense checking), and avoid clicking unknown links.

Passive hijacking attack. The attacker in passive hijacking monitors the network traffic to capture the authenticated information and does not interfere with the data flow or storage. These types of attacks are hard to capture or discover. This is further categorized into:

- *Network Traffic Monitoring:* It is used to understand what is moving in and out from the network to detect or monitor an event of interest. Earlier it was used for network anomalies detection, but malicious users tried to analyze the traffic to intercept the message for data confidentiality and data integrity breach.
  Defense methods: Connect with only trusted network, encrypt the network traffic flow, and continuous network monitoring or scanning for sniffers.
- *Network Sniffing:* In this type of passive hijacking attack, an attacker monitors or intercepts the data flowing over the network communication channels in real-time. Network sniffers use sniffing tools such as Wireshark, CloudShark, Microsoft Message Analyzer, Omnipeek, and Ettercap to captures the data copies without redirecting or altering it. An attacker able to get the data analysis report and intercept it to break the confidentiality and integrity of data.
  Defense methods: Connect with only trusted network, encrypt the traffic flow, and continuous network monitoring or scanning.

4.3.2. Communication interruption attack

In this type of attack, the aim of the attacker is to obstruct the network service for authorized users and leaves the system unusable. This attack is against the availability of the system. Examples of interruption attacks are server overloading, disconnecting communication lines, and redirect requests to an invalid address. Various types of communication interruption attacks are shown below.

Denial of service attack. This attack makes the network resources unavailable to the legitimate users by temporarily or indefinitely obstructing the services of host or web server. It can be accomplished with flooding, smurf, SYN, and buffer overflow. Attackers may hamper the data processing and analytics procedure by disrupting either system resources or communication links.

Defense methods: Deploy antivirus and firewall into the network, contact ISP immediately, and Investigate black hole routing.
Various possible DoS attacks are:

- *Ping Flood Attack:* It is also known as the ICMP flood attack, which attempts to saturate the target system with echo-request packets to hamper the resources required to process and analyze the data. Ping command is generally used to check the connectivity between the systems with echo-request packets.
  Defense methods: Reduce the TTL of the ping packet, install firewalls and antivirus.
- *SYN Attack:* It is a type of DoS attack which continuously sends SYN requests to the target system to consume server resources. It makes the server non-responsive to the authorized users so that data analysis and processing gets affected and incomplete.
  Defense methods: Message filtering, reducing SYN packet TTL, and configure proxies and firewalls.
- *Buffer Overflow Attack:* It is an anomaly program that writes an anomaly data into a buffer that overflows the buffer memory. It is generally triggered with malformed inputs and affects the data analysis procedure with incorrect data inputs.
  Defense methods: Choose an appropriate programming language, use safe libraries, and pointer protection.

Distributed denial of service attack. It is the next level of DoS attack in which incoming traffic is flooded from different heterogeneous sources to make it impossible to stop. Since 2016 the scale of DDoS attack is increasing day by day. DDoS attackers target the application layer processes to disable resources and services.

Defense methods: Increased bandwidth, contact the ISP as soon as possible, Create a DDoS playbook, and configure antivirus and firewall

- *HTTP Flood Attack:* It is a volumetric DDoS attack from different sources to overwhelm the targeted host. It is of two types such as HTTP GET and POST attacks.
  Defense methods: Authorize the IP range, authenticate the web server, and restrict a number of parallel connections.
- *Smurf Attack:* In this type of attack, a large number of ICMP packets are broadcasted over the victims network with spoofed source IP addresses. This may slow down the victims system.
  Defense methods: Configure hosts not to respond to ICMP messages and not forward broadcasted packets.
- *UDP Flood Attack:* It is an attack with a large number of UDP packets that are broadcasted to the intended target with the aim of obstructing the services and resources. This attack exploits the server by targeting one of its ports for UDP packet flood.
  Defense methods: Maximize bandwidth, apply message filters, and configure firewalls.

GPS spoofing attack. This attack aims to mislead or redirect the GPS by broadcasting fake or malicious GPS signals. Attackers modify the spoofed signals in such a way which shows the real location somewhere else. It has the power to redirect the confidential network traffic data to their specified location and manipulate it.

Defense methods: Non-locatable antennas, install sensor or blocker and reduce network latency.

> This section discussed the possible security attacks on data analytics in detail as per the taxonomy mentioned in Section 2. The possible defense mechanisms is also mentioned at the end of each attack explanation.

5. Secure data analytics architecture and threat model

A threat model is to design policies against various types of security threats and possible mitigation strategies. A threat model can answer the following questions: (i) What are we designing? (ii) Is the system under security threats? (iii) What should we do against the security threats? and (iv) Are we performing stunning job? The detailed architecture for SDA and the threat model processing is explained in the following subsections.

5.1. Threat model

Most of the common information and software security problems are due to complex solutions, unverified assumptions, or dependencies on external entities. To protect the system, we need to explore all possibilities before designing any solution for attack mitigation. To handle these problems, we need to evaluate the system with threat modeling in a security context. Moreover, we need to analyze the system from attackers perspective and need to identify the ways which breach

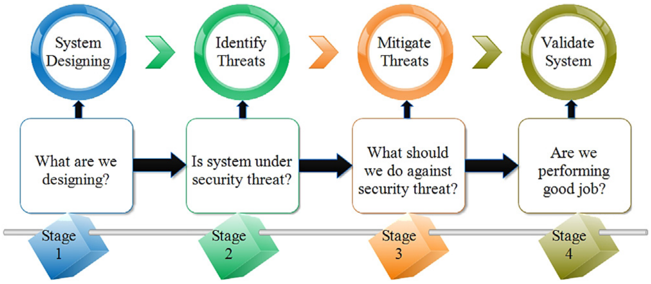


Fig. 6. Stages of threat model.

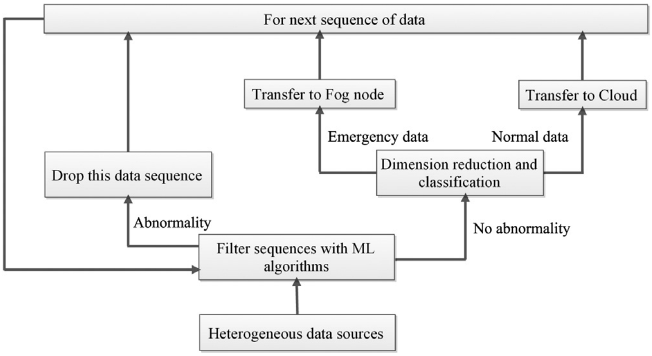

For next sequence of data
Transfer to Fog node
Transfer to Cloud
Emergency data
Normal data
Drop this data sequence
Dimension reduction and classification
Abnormality
No abnormality
Filter sequences with ML algorithms
Heterogeneous data sources

Fig. 7. ML-based threat model for SDA.

the security properties like availability, integrity, and confidentiality. Different threat modeling methodologies are available in the literature depending upon the volume of data such as-(i) STRIDE modeling, (ii) Process for Attack Simulation and Threat Analysis (PASTA), (iii) Trike modeling, (iv) Visual, Agile, and Simple Threat modeling (VAST), and (v) OCTAVE modeling. The resultant threat model enumerates all attacks and produces the traceability matrix, which ranks the risk associated with various types of attacks. We need to look at these threats and check it's mitigation strategies. If not, then look for its potential design.

The various stages of the threat modeling is as shown in the Fig. 6. Methodologies to mitigate generic threats (store and transfer data securely) are available such as-cryptography, but mitigation of threats vulnerable to data analysis still needs to be explored. Cryptographic techniques for data security are not efficient because they can identify only one or very few numbers of attacks or threats. If any other type of attack is encountered in the system, then cryptographic algorithms are not adequate to identify it. So, we need to have a secure system that can learn the patterns of the existing threats and predict a similar threat based on the leaning. To achieve it, ML or DL models can be a viable solution that can overcome the issues in identifying similar or new threats. ML has the capability to learn the hidden patterns of threat using back-propagation to identify similar data analysis security threats. The proposed threat model for SDA is as shown in Fig. 7.

The threat model takes the input from the heterogeneous data sources such as sensors, environment, and digital libraries and apply the ML algorithms to filter out the input sequences in terms of abnormality or not. The preferred ML algorithms used in this stage of threat modeling are LSTM and decision tree. LSTM predicts the behavior of data stream coming from one single source based on their previous data patterns, whereas the decision tree classifies the output of LSTM into suspicious or normal. If it is suspicious, then drop that data pattern and read the new one. Otherwise, feed the data into a principal component analyzer for dimension reduction (as it is big data) for better processing [29]. Once the data dimensions are reduced, it is then fed as an input to the decision tree classifier. Here the role of decision classifier is to bifurcate the requests based on the criticality of data (request to ambulance, police, doctor, etc are critical requests). The critical requests are forwarded to the fog nodes (fast processing) and others are to cloud server. After the processing of the previous request, the new data sequence will be processed.

The objectives of a threat model is to verify the security issues with system design, apprehend threats, basis for secure system implementation, and reliable system design. The existing threat models are using cryptographic-based methodologies such as-Homomorphic Encryption Method (HEM), Differential Privacy Method (DPM), and Pseudonym Certificate Technology (PCT). HEM technique performs random data computation on cipher-text, which generates an encrypted result. It is used to secure cloud computing services without revealing sensitive data such as-tax information, voting data, and health data. HEM is categorized into Full HEM (FHEM) and Partial HEM (PHEM). FHEM supports mixed data computation, whereas PHEM supports restricted application scenarios. DPM is a privacy-preserving technique that adds an impulsive noise in the user data. In PCT, user requests cloud services using pseudonym certificates. It protects users location and identity privacy. Its security score is lower as compared to the other discussed techniques.

A relative comparison of the ML-based threat model with other state-of-the-art existing threat models is shown in Table 5.

5.2. *SDA architecture*

This section discusses both traditional and ML-based SDA architectures. The detailed description of both the architectures is explained in the following sub-sections.

5.2.1. *Traditional SDA architecture*

The purpose of SDA architecture is to secure the data analysis. The data is to be collected from different heterogeneous repositories (*machines, Internet, Image data, video data, and enterprise*) and external environments (*agriculture sensing data, smart grid data, border areas, gas plants, and water covered areas*). Traditional SDA architecture uses cryptographic methods to identify, prevent, and mitigate security threats. Such cryptographic techniques are HEM, DPM, and PCT. These methods are likely to identify only one or a few threats. If any similar kind of threat tries to breach the security, then cryptographic techniques are not adequate to handle it because of their limited scope of existing security algorithms.

The traditional SDA architecture is outlined in Fig. 8, which has six components (i) data sources, (ii) sensors, (iii) cloud, (iv) fog nodes, (v) end nodes, and (vi) communication channel. Description of these components are as follows:

- *Data Sources*: It is a repository in which data is being captured for computation and further analysis. Data sources or sites can be either homogeneous or heterogeneous. In similar data sources, all sites are using the same DBMS product, whereas heterogeneous data sources sites use different DBMS products. The other way to capture data is from the sensors, which gives real-time data, whereas the repository provides stored old data. Sensors are used to collect data from locations in which human involvement is at high risk such as-chemical plant, forest monitoring, and gas plant.
- *Sensors*: Sensors are lightweight and low-power devices used to capture real-time data from the external environment. They can be classified as Accelerometers, Biosensors, Image Sensors, and Motion Detectors based on application areas. Moreover, they collect real-time data and store it into the centralized cloud infrastructure for further use.
- *Cloud*: It refers to access to the software, platform, and infrastructure services through the Internet. It stores the data at a centralized location. Users can store, process, and manage the data on the cloud instead of a personal computer (called distributed data processing). Data can be accessed anywhere and

Table 5
A relative comparison of the proposed ML-based threat model with other state-of-the-art existing threat models.

	OCTAVE	Trike	P.A.S.T.A	STRIDE	VAST	ML-based
Implement application security at design time	✓	✓	✓	✓	✓	✓
Identify relevant mitigating control	✓	✓	✓	✓	✓	✓
Prioritize threat mitigation efforts	✓	✓	✓	✕	✓	✓
Encourage collaboration among all stakeholders	✓	✓	✕	✕	✓	✓
Automation for threat modeling process	✕	✓	✓	✓	✓	✓
Integrates into agile DevOps environment	✕	✕	✕	✕	✓	✓
Ability to scale across many threat models	✕	✕	✕	✕	✓	✓
Ability to predict the novel threat	✕	✕	✕	✕	✕	✓
Ability to learn and design new mitigation techniques	✕	✕	✕	✕	✕	✓
Ability to identify multiple threats	✕	✕	✕	✕	✕	✓

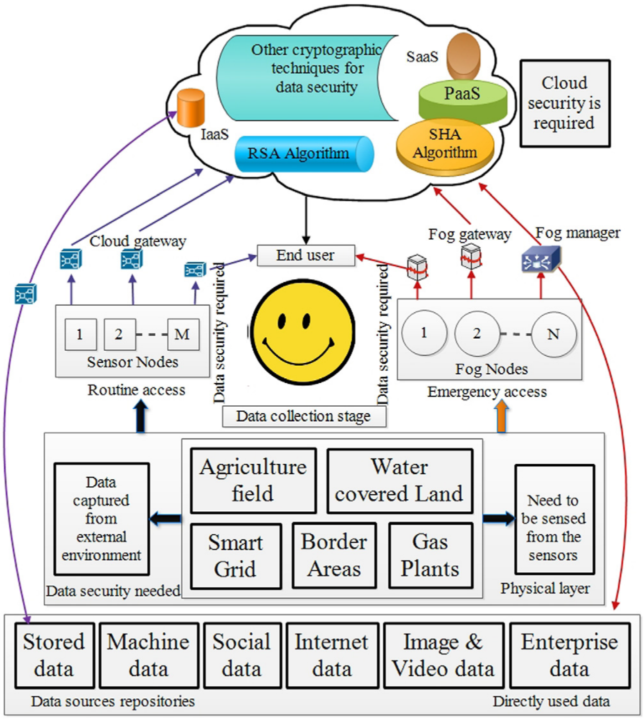
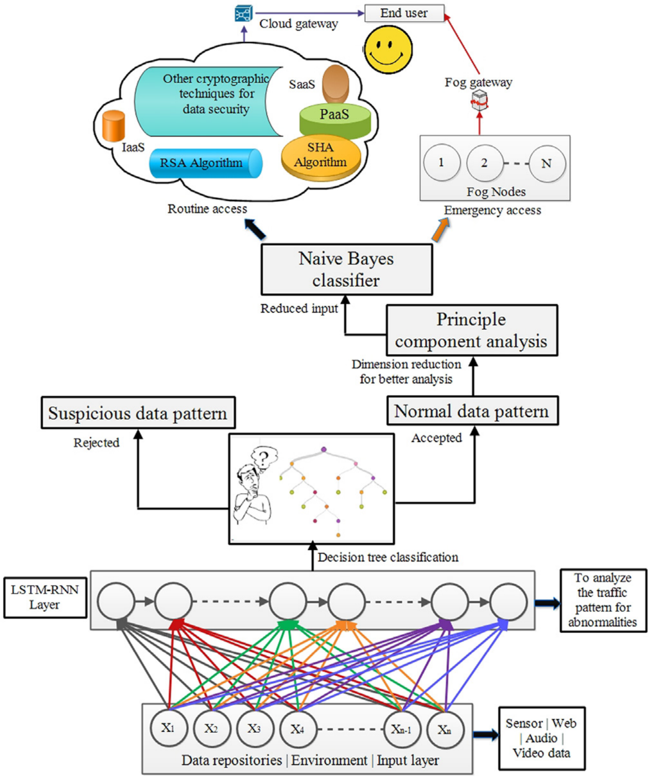


Fig. 8. Traditional SDA architecture.



Cloud gateway
End user
Other cryptographic techniques for data security
SaaS
PaaS
SHA Algorithm
IaaS
RSA Algorithm
Routine access
Fog gateway
1
2
N
Fog Nodes
Emergency access
Naive Bayes classifier
Reduced input
Principle component analysis
Dimension reduction for better analysis
Suspicious data pattern
Normal data pattern
Rejected
Accepted
Decision tree classification
LSTM-RNN Layer
To analyze the traffic pattern for abnormalities
X1
X2
X3
X4
Xn-1
Xn
Data repositories | Environment | Input layer
Sensor | Web | Audio | Video data

Fig. 9. Machine learning-based SDA architecture.

anytime. Data security is provided with the use of cryptographic algorithms such as HEM, DPM, and PCT, whereas security to data computation and analysis can be compromised.

- *Fog Nodes*: They are like a mini-cloud installed at the edge of the network. In fog computing, the storage, services, and computations are close to the end devices, which provides faster access to data. But, the problem with fog computing is its low computing power compared to cloud computing. It offers more security to the data as it is near to the user's proximity. But, the security related to the data analysis needs to be addressed.
- *End Nodes*: These are the peripheral devices operated by users. It can be a personal computer or mobile device.
- *Communication Channel*: It is a network channel for communication among the various components. Sensors sense data and transferred to the centralized cloud through a communication channel for further processing.

5.2.2. ML-based SDA architecture

Traditional SDA architecture is inadequate at the data analysis phase such as-not able to identify threats at the time of data analysis. For example, if the sensor reads a specific real-time data such as-motion, bio, and health-related data (blood pressure, heart rate, and pulse rate) from the external environment, then there may be a chance that intruders can attach their data packets with the real-time data to make sensors reading wrong. This can be mitigated with the usage of ML models during the data reading phase. ML-based SDA architecture is shown in Fig. 9.

A Long Short Term Memory (LSTM) Recurrent neural network (RNN) can be used in ML-based SDA architecture to provide security against the sensor data reading. It learns and recalls the input data patterns at time T to identify the irrelevant data patterns from the input network traffic. It has the capability to remember long patterns and predicts the next word in sequence. After reading the correct data by the sensors from the external environment using ML, LSTM based Decision Tree (DT) classification algorithm can be applied to classify the type of input traffic data into-suspicious or normal. If the input data is classified as suspicious, then it is removed from the input sequence. Otherwise, it is taken as an input to the Principal Component Analysis (PCA) model. It reduces the dimensions of the data without losing its features for better analysis and processing.

The output of the PCA model is in the reduced form having different processing priority levels. Data priority means how important the data is? If data priority is high (healthcare data, calling an ambulance, and suspicious activity) and is frequently used, then it can be kept on fog nodes/gateways for quick access and analysis. Otherwise, it can be stored in the cloud (entertainment data, inventory data, and census data). It is because the data fetching time from the fog is lesser as compared to the cloud. ML-based SDA architecture can use the Naive Bayes classifier to classify the data into high or low priority levels. Detailed SDA architecture with ML algorithm is as shown as in the Fig. 9.

ML-based SDA architecture can provide more security during the data analytics as compared to the traditional SDA architecture. Moreover, it can learn the hidden patterns in the network communication sequence with RNN and may use the classification algorithms to block the unusual traffic patterns generation sources. It helps to detect and mitigate multiple types of attacks (DoS, U2R, R2L, probe, DDoS, and zero-day attack) instead of one as in traditional cryptographic algorithms.

This section presented the various stages of threat model and the ML-based SDA threat model to predict and classify the data stream as specious or attack free. It also presented the ML-based SDA architecture in comparison to the traditional SDA architecture. This section also highlighted the comparison proposed threat model with the traditional ones.

6. Case study: Smart grid

To validate the process of the proposed architecture, we presented a case study on the smart grid (SG) system. The SG generates from the different sources (wind power plant, hydropower plant, solar power plant, and many more) and distributes it to the end-users. The consumption of energy is recorded in the SG system through smart meters and sensors [30]. The analytics on energy consumption is beneficial for the SG system to distribute energy as per the user requirement in an efficient way. The traditional approach for data analytics may collect un-secure data from the data generation sources and may result in inaccurate decision making. Hence, there is a need of SDA for accurate decision-making [31].

In this paper, we proposed an ML-based SDA architecture to predict the data traffic patterns using the LSTM model for security threats and classify them accordingly. Initially, the energy is collected from smart meters and fed as an input to the LSTM layer for traffic pattern analysis. The analysis output of the LSTM is connected to the decision tree classifier, where the energy data is classified as suspicious or normal. The next phase reduces the dimensions using PCA to store secured data on the cloud/fog platform [32]. Now, the analytics will be performed on the filtered secured energy data and gives an accurate result.

7. Taxonomy of machine learning-based secure data analytics

The detailed taxonomy of ML-based SDA is summarized in Fig. 10. The detailed description of each existing approach is discussed in the following subsections.

7.1. Distributed processing-based SDA

The volume of the data is exponentially increasing day-by-day, and it is not possible to store it at one single site, which in turn decreases the efficiency of system such as-throughput and delay. One of the best ways to store such a massive volume of data is at distributed locations [33]. Various possible techniques to store the data at distributed sites are discussed in the subsequent paragraphs.

7.1.1. Cloud-assisted SDA

Cloud Computing (CC) is an emerging technology, which stores numerous amount of data at remote servers. It allows anytime and anywhere access to the data. In CC, the data is online, so it is more vulnerable to attacks such as-confidentiality and privacy attacks [34]. Thus, financial firms are moving towards the development of Cybersecurity Insurance (CSI) procedures to protect their data. The cost of CSI procedures is too high. So, Gai et al. [35] developed a framework to reduce the CSI cost without compromising the security level of BD.

Bothe et al. [36] presented an eSkyline query interface to process skyline queries on an encrypted data. It answers the queries posted by clients without disclosure of actual data values based on the Attribute-Order-Preserving-Free-SFS (AOPFS) algorithm. It provides security to cloud-based databases. However, the problem with this system is that it is not suitable for large datasets. Later, Cuzzocrea [37] extended the work presented in [36] to perform secure BDA based on the same AOPFS algorithm. The author used a skyline operator to perform secure BDA, which becomes effective and reliable using the AOPFS algorithm. It was focused on BD compression with reference to immense data IoT frameworks. It is currently known as an emerging interdisciplinary field that incorporates service-oriented infrastructures, CC, BD management, and analytics. Puthal et al. [38] proposed a security verification model that considers the security aspects of BDA and provides real-time analysis of CC services. Moreover, it analyzes the data stream and generates an alert for any emergency event which supports end-to-end security to prevent data stream from unauthorized access. It provides security to both data and query with proposed dynamic prime number-based security verification and dynamic key length based security frameworks with the shared key. The authors have considered three levels of security, strong, partial, and no confidentiality levels. The main focus of this architecture was to generate data alert in an emergency to block unauthorized access. The data shared on the cloud was not protected at all and can be misused by the intruder. Later, Ojha et al. [39] presented a cryptographic-based technique to protect data from the unauthorized users. They used the Advanced Encryption Standard (AES) and Message Digest (MD5) algorithm to make the data more secure. It encrypts the login id and password (thumb impression) of the database to protect against a man-in-the-middle attack. But, it does not consider any other data security breach except man-in-the-middle attack, DoS attack, and replay attack.

Murali et al. [40] proposed an architecture that provides security to the data shared on the cloud using quantum cryptography. It has three phases — registration phase, communication phase, and an authentication phase. This architecture authenticates the end-users. They tested the architecture with BB84 protocol and simulated using the QKD simulation environment. The experimental results were evaluated with eavesdropping probability between 0.1–0.9 with 500 qubits each. They claimed that their architecture provides a high level of confidentiality and authentication. In [41], authors developed a system that guarantees the correctness of the data, which creates trust among users to store their data to the cloud server. The delay in processing the data by cloud-assisted SDA is high, which is not suitable for latency specific applications such as telehealth, telemedicine, and the autonomous car. The solution to this issue is elaborated in subsequent sections.

7.1.2. Fog-assisted SDA

Fog Computing (FC) increases the efficiency and reliability of the cloud system by keeping fog nodes near to the proximity of the end-users. It also reduces the delay in accessing data from the cloud. It stores a specific small chunk of data in it [42]. Authors in [43] presented a fog-based framework *(FogGIS)* for performing analytics from geospatial servers. FC increases the throughput, reduces the delay, and reduces the transmission power. Fog devices act as a low power gateway between clients at different locations, which used the Intel Edison processor for computing [44]. This system reduces the delay and the transmission power, which increases the throughput of the

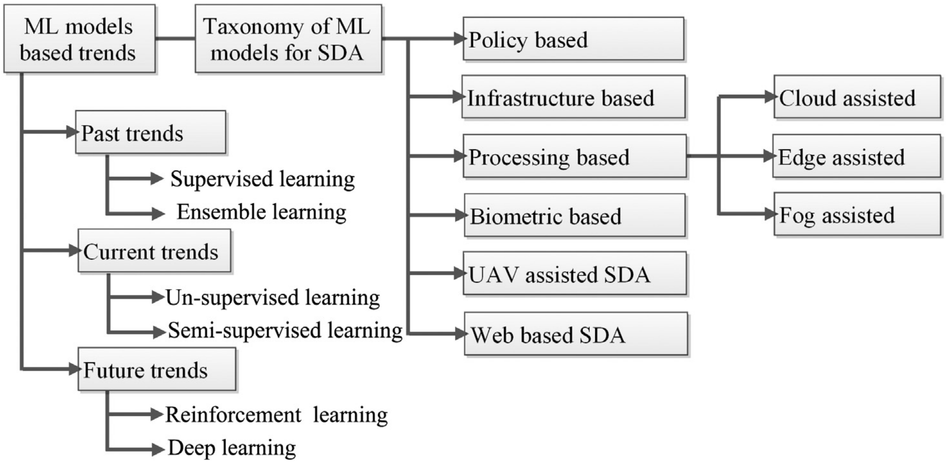


Fig. 10. Taxonomy of ML-based SDA.

system. Irrespective of geographical data, FC is useful for other applications such as-smart healthcare [45], smart agriculture, and smart logistics [46].

With technological advancements, smart devices such as-mobiles, wearables, and tablets are capable of generating a large amount of data, i.e., every millisecond calculation [47]. Analyzing such a massive amount of data with traditional database techniques and algorithms is a challenging task. So, BDA has the capability to process a tremendous amount of data. Existing CC solutions are not viable to handle a large amount of data in real-time because of more delay and transmission power. Mehdipour et al. [48] addressed these issues in their proposed framework known as *FOG-engine*, which performs analytics on a large amount of data before it was uploaded to the central location. This *FOG-engine* increases the efficiency in terms of processing speed, network bandwidth, and data transfer size. Moreover, it was directly managed by end-users and allowed data to be processed at the edge of the connected IoT devices. It has solved the issues of throughput and latency, but still, the problem of cost and data security persists.

Hernandez et al. [49] proposed a 3-tier architectural framework for real-time data processing based on edge-fog technology. But, the security of data was still an issue in this 3-tier architecture. Dsouza et al. [50] proposed a secure policy-based resource management in FC. The feasibility of the proposed system was evaluated with a proof-of-concept, but the data was still not fully protected. Later on, Dang et al. [51] addressed the data security issues in FC and proposed architecture with three different models such as-region-based, privacy-based, and mobility-based. They achieved better results with their proposed system compared to the cloud data access and security.

7.1.3. Edge-assisted SDA

In the current era, every device needs to be connected with the Internet to share and receive information from the far-flung nodes. The cloud and fog nodes provide easy access to the data anytime from anywhere. The latency and throughput of data access in FC are lower than the CC [52] because the fog nodes are located near to the client machines. However, the latency and throughput of the system can be increased with the usage of mobile MEC. In MEC, the processing is at the client's device, not at the remote site (cloud and fog). So, The security of data in MEC is still an important issue. Cui et al. [53] proposed a proxy-aided attribute-based encryption standard for data security. This approach performed decryption at edge devices and does not need any secure channel for key distribution. Its performance was evaluated in charm framework with python and PBC library, which reduces the computational overheads significantly. But, this approach is suitable for a small volume of data, and if the data volume increases, it's performance decreases. Zhou et al. in [54] presented a trustworthy multimedia BDA approach for edge computing for social multimedia data analysis.

Table 6 shows the relative comparison of existing state-of-the-art distributed processing techniques concerning parameters such as-methodology, tools, pros, and cons.

7.2. Policy-based SDA

A few decades ago, small databases were sufficient to store and process the data. Now, the world is digitizing, which leads the automation of tasks such as-machines can perform human capable tasks. Digitization increases the data with a rapid rate, which needs more attention in context with secure access from different locations. For example, sensors are capable of capturing data every second, which arises the need for large repositories like data warehouse for managing the BD. These repositories are capable of storing and handling an enormous amount of data. So, the data warehouse is a repository to store extensive data, whereas BD is a technology that manages a large volume of data efficiently. But, these repositories opens the door for attackers to breach the security of the stored data because of the poor traditional infrastructure. So, the protection of data against various threats is now a primary concern for every organization. To handle and secure such a significant amount of data, we need standardized policies and frameworks. Data policies are high-level statements which show how data is to be managed.

Many countries throughout the globe have framed country-wide specific policies for personal data security. Recently in 2018, National eHealth Authority (NeHA), India has presented a policy draft named "*Digital Information Security in Healthcare Act (DISHA)*" to protect patients privacy data such as-e-health records [57]. This act contains policies and framework to protect the personal data of Indian users. This act mandates "*data localization*", i.e., all data generated by an Indian payment system is to be stored in India only [58]. Healthcare organizations of various countries devised their country-specific policies based on their data security requirements.

The first e-health data security policy was the Health Insurance Portability and Accountability Act (HIPAA) framed in 1996 by the US Congress. The HIPPA act specified both privacy and security guidelines to secure patients health records of the US. Later on, different standards such as-Medicare, PPACA, PSQIA, and HITECH act were developed by the USA health and welfare department to protect their citizen's healthcare-related data [59,60]. Likewise, other countries also framed policies to secure the patients personal e-health records. The detailed description of country-wise healthcare standards are mentioned in Table 7.

Likewise, security and privacy policies for other sectors were also framed by their respective organizations. In agri-business "*Latevo Farm*

Table 6
Relative comparison of state-of-the-art distributed processing techniques.

Author	Year	Objective	Methodology	Pros	Cons	Tools
Dsouza et al. [50]	2014	Presented the resource management policy for fog computing to secure and process user requests	Descriptive	Support interoperability between resources and proved the feasibility with proof-of-concept	Policy conflicts are difficult to resolve and effectiveness of different devices interoperability not measured	OpenAZ
Bothe et al. [36]	2014	Devised a eSkyline-based framework to process encrypted queries	Descriptive and algorithmic	Query on encrypted dataactual data values not disclosed	Not suitable for big datasets and distributed data	–
Cuzzocrea [37]	2016	Presented a reference architecture secure BDA over Cloud Databases	Experimental	Secure BDA for cloud databases	Not for real-time data security	Map Reduce Hadoop
Puthal et al. [38]	2016	Proposed a framework for event detection in data stream to avoid unauthorized access and manipulation	Analytical	Secure BD stream	Not Implemented	Apache Storm tool (also called as real-time hadoop)
Gai et al. [35]	2016	Presented a BD cloud-based cybersecurity analysis framework	Experimental	Cybersecurity on BDA	Better feasibility and adaptability	business strategic and analytical tools
Barik et al. [43]	2016	Proposed a fog-based framework to analyze and validate geospatial data	Analytical	Geospatial data analytics, reduces latency, reduces storage, and increases efficiency	Not suitable in mobile client environments	GIS tool
Mehdipour et al. [48]	2016	Designed a fog engine for IoT devices to facilitate data analysis before loading it into a centralized location	Descriptive	Low latency, high throughput, and low bandwidth usage	Low security, high cost and low extensibility	IoT device setup
Hernandez et al. [49]	2017	Presented an architecture to establish real-time stream data flow for immediate response	Experimental	Filter the redundant data	Not provided temporary storage at the edge and not for heterogeneous platforms	Kinetic Cisco platform
Murali et al. [40]	2017	Designed a secured protocol for cloud authentication with the use of quantum cryptography technique	Experimental	High confidentiality and authenticity for data security and reduces key extraction time	Costly approach with low scalability	AVISPA tool and QKD simulator
Ojha et al. [39]	2017	Authors used AES and MD5 cryptography approaches to encrypt and decrypt data at login time only without authentication	Conceptual	Easy to implement, low complexity, less costly	Authentication process not implemented	Dot NET environment
Dang et al. [51]	2017	Proposed a fog-based privacy control method to handle device location changes.	Experimental	Protects data and handles mobility	Not includes feasibility and efficiency of the model.	Mobile-based testbed
Cui et al. [53]	2018	Designed a proxy-based cypher-text policy to secure and decrypt edge device computations without the need of any secure channel for key exchange	Analytical	Solved latency, security and privacy issues	Unknown security threats on key exchange is not highlighted	Charm python framework
Zhou et al. [54]	2018	Presented a trust-based contextual learning algorithm for social media data analysis	Experimental	Data privacy and suitable for big datasets	Accuracy based on user preferences	Data analysis tools
Garg et al. [55]	2019	Presented a Security Framework using edge computing for big data analytics in vehicular ad hoc networks	Experimental and algorithmic	Energy efficiency and low latency	–	Quotient
Alabdulatif et al. [56]	2020	Given the privacy-preserving framework for big data in cloud environment using homomorphic encryption	Mathematical modeling and algorithmic	Increases the performance and analysis accuracy	Homomorphic encryption is slow	Homomorphic encryption

Table 7
Comparison of country-wide data protection standards.

Policy name	Regulatory body	Country	Year created	Year en-force	Description	Application areas	Pros	Cons
Constitution	Brazil and the federal government of Brazil	Brazil	1988	1988	It improves private life, intimacy, honor, and image of people.	Personal and Health data	Regulate labor relations and data security	More possibility of security attacks
Data Protection Directive Regulation (DPR)	Information Commissioner's Office (ICO)	Europe	1995	1998	It is also known as Directive 95/46/EC and used to process the personal data within European Union (EU).	Personal Data	Its seven principles such as notice, purpose, consent, security, disclosure, access, and accountability provides more data protection.	Strictly binded with EU members.
Personal Information Protection and Electronic Documents Act (PIPEDA)	Innovation, Science and Economic Development Canada (ISED)	Canada	1995	2000	It is privacy policy to protect and governs private sector data.	Private sector data	Accurate, latest and complete information, it includes human anti-discrimination	Works well only internally without risk
Health Insurance Portability and accountability Act (HIPAA)	United States Department of Health and Human Services	USA	1996	2003	It is a national standard to secure e-health records.	Health data	Protect security and privacy, Penalty on confidentiality violating	Not for individuals below age 12 years, lack of trust between patients and doctors, lack of patient control
Data Protection Act (DPA) 1998	Information Commissioner's Office (ICO)	UK	1998	1998	Used to safeguard business and personal data stores on personal computers.	Business and Personal Data	Effective business management, customer security	Complex interpretation, limited to European Economic Area, high cost, and training
IT Act and IT (Amendment) Act	Indian Computer Emergency Response Team (CERT-In)	India	2000	2000	It protects against cybercrime and civil rights of an individual by providing legal transaction recognitions.	Personal Data and Civil Rights	Reduces cybercrime, validates eSignatures and contracts	More vulnerable to attacks
Russian Federal Law on Personal Data	Federal Law Russia	Russia	2006	2009	It protects rights and freedom of human and their personal data.	Personal Data	Avoids personal data destruction, provides data confidentiality	Not cover relations during data processing
Patient Safety and Quality Improvement Act (PSQIA)	Agency for Healthcare Research and Quality (AHRQ)	USA	2008	2009	It improves the quality of medical services and safety of patients data.	Health data	Medical services with highest standard of care, no need of approval	More risk is associated
Health Information Technology for Economic and clinical Health Act (HITECH Act)	United States Department of Health and Human Services	USA	2009	2009	It is used to promote eHealth adoption	Health data	Improved quality, safety and efficiency, increase care coordination	Chances of attacks increases
The 09-08 act, dated on 18 February 2009 [61]	National Data Protection Authority (Data Protection National Commission)	Morocco	2009	2009	It protects personal data, sensitive personal data, and its implementation.	Both personal and sensitive personal data	Encourage foreign investment, authorization before data transfer, accurate and up-to-date	Data security during abroad transfer and required data retention procedures
Patient Protection and Affordable Care Act (PPACA)	Centers for Medicare & Medicaid Services	USA	2010	2014	It is also called as Obamacare or Affordable Care Act. It includes both healthcare and education sectors.	Health data	Healthcare cost reduces, Education sector covered	Health insurance companies discontinue plans of many people
General Data Protection Directive Regulation (GDPR) [62]	Information Commissioner's Office (ICO)	Europe	2016	2018	It is an enhanced version of DPR. It provides security against privacy and data breaches.	Personal Data	Increased scope of privacy, consent conditions strengthened, and data portability	Mandates to inform before 72 h on data breach
DISHA	National eHealth Authority (NeHA)	India	2017	2018	Used to protect patients privacy data and regulations.	Health data	Accurate diagnostics, data confidentiality, privacy preserving.	Ignores circumstances in which health services availed.

Income Protection" was developed for farmers income assistant [63]. Comparison of various country-wide standards for private and public data security and privacy, as explained in Table 7.

7.3. Machine learning models-based SDA

We are now moving towards the era of Industry 4.0, i.e., digitization, where the focus is to keep everything connected and online 24 by 7. But, anything connected to the Internet is more vulnerable to security attacks. Traditional security algorithms such as-RSA, DES, ECC, and PKI are available to handle the aforementioned attacks. But, these algorithms can detect only those attacks which are in their knowledge base [12]. Moreover, nowadays many new attacks are influencing the system and breaches their security, which makes the existing algorithms inadequate to identify such new types of attacks. This issue can be resolved with the help of ML models or algorithms [64]. Various ML models are used to handle such attacks are shown in Fig. 11, where ML models are categorized into past, current, and future trends based on their usage and applicability.

7.3.1. Past-trends

This section highlights the ML techniques or models which were exhaustively used earlier, but outdated in the present era.

Supervised learning models (classification models). Classification models apply to the only situation where the class labels of test data are given. It is a two-step process such as-(i) *learning step* and (ii) *classification step*. In the learning step, the model is trained and is called as the classifier and is used to predict the class label of the new data in the classification step. There are various classification-based models:
Logic-Based Techniques:

- *Decision Tree:* Security risks to the network increase as the network-services increases. Security is a significant concern nowadays for any individual in the world. Many systems are available to identify and mitigate security attacks such as-intrusion detection systems (IDS) and anti-viruses. These systems are liable to detect and prevent only those attacks which are in their security database and fail to recognize new security threats. Various limitations with the existing systems are: (i) limited attacks detection, (ii) not possible to specify all intrusions, and (iii) encoding rules process is costly and slow. To overcome these limitations authors in [65] proposed an algorithm which used alternating decision tree technique in combination with well-known ensemble technique called a "*boosting*" to mitigate IDS attacks such as-probe, DoS, R2L, and U2R attacks with feature selection. They have used the NSL-KDD dataset for the experimental analysis on Weka open-source tool to estimate the parameters such as-(i) accuracy, (ii) detection rate, and (iii) false alarm rate. Its performance was quite better than the existing naive Bayes model, and it also classifies attacks effectively.
  Hanmanthu et al. [66] proposed a methodology based on classification technique to prevent SQL injection attack. The decision tree was used to identify the class of HTTP requests. If the request is an attack, then it provides a negative response; otherwise, a positive response. Other models are also available to detect SQL injection attacks such as-Netsparker, WebCruiser, and Acunetics. Authors have evaluated the performance of the existing non-ML-based approaches with the decision tree-based approach based on the parameters such as-accuracy and attack detection time. They achieved 82% of accuracy which is better than the other approaches and attack detection time is quite less compared to Netsparker approach, but higher than other approaches.
  Later, Lakshminarasimman et al. [67] proposed an algorithm to detect DoS and DDoS attacks using decision tree supervised learning technique to make the wireless communication secure. It finds the attack pattern and provides a suitable counter mitigation steps. Experimental analysis was performed with the KDDCup'99 dataset.
- *Rule-Based Model:* It is a classification technique that classifies a large amount of data into different classes based on the production rules formulated from the given labeled test data. It is used in network security to analyze the network traffic for intrusion detection. Komviriyayut et al. [68] analyzed both decision tree and rule-based classification techniques for intrusion detection with online dataset RLD09. They classified the data as normal and attacked data (DoS and probe attack). They concluded that the intrusion detection over the data having more features (*RLD09 has 13, and KDD99 has 41 features*) with decision tree produced better detection rate over the rule-based technique. Authors of [69] used five different rule-based supervised learning algorithms to detect network intruders such as-(i) Decision Table, (ii) JRip, (iii) OneR, (iv) PART, and (v) ZeroR. The decision table uses tabular representation, JRip optimizes rule set to add each possible rule, OneR for discrete data values, PART encapsulates both divide and conquer techniques to design rules, and ZeroR is target based which has no predictability power. The experimental analysis was done on the WEKA tool with a cross-validation and KDD-CUP dataset. In both cases, authors found that PART rule-based algorithm gives a better percentage to classify data correctly.
  The standard rule-based algorithms are not able to classify the data accurately. Xue et al. [70] designed an algorithm to generate an attack signatures accurately. They experimented with their algorithm over KDD CUP 99 dataset Apache-Knacker, BIND-TSIG, and ATPhttpd attacks. They concluded their algorithm is better compared to the other two analyzed algorithms such as-Token Subsequence and ISW.

Statistical Learning: It is an ML technique based on conditional probability to identify the class label of the unknown instance. In this paper, we discuss two types of statistical learning techniques such as-Naive Bayes and Bayesian Network Classifiers. The implementation of these techniques in network security is described below.

- *Naive Bayes Classifier:* It is probabilistic classifier based on *Bayes Theorem*. In this classifier, every feature being classified is independent of each other. It computes the conditional probability of each row in the training dataset and predicts the class label of the new tuple. Similar to logic-based techniques, the Naive Bayes model is also used to detect network security threats. Deshmukh et al. [71] proposed a system that uses the existing Naive Bayesian classifier with discretization, normalization, and feature selection methods. Their system improved the results of the Naive Bayes classifier in terms of time taken in anomaly detection. They compared three classification algorithms such as-NB Tree, AD Tree, and Naive Bayes. They have used NSL-KDD 99 dataset with four types of threats such as-(i) DoS, (ii) probe attack, (iii) remote-to-local (R2L), and (iv) user-to-root (U2R). They concluded that the time taken to build a model with Naive Bayes is less compared to the other two considered algorithm.
  Yang et al. [72] modified the existing Naive Bayes algorithm using an artificial bee colony algorithm to detect network anomalies. Artificial bee colony algorithm is a heuristic algorithm inspired by honey collecting strategy of bees, which gives higher and lower weights to the neurons based upon their greater or smaller impact on results, respectively. They have tested the algorithm over Sonar, Biodeg, and NSLKDD datasets with 5000 instances for its accuracy to detect network intruders. The accuracy of the modified Naive Bayes algorithm (>90%) is better than the traditional Naive Bayes (70%–80%) classifier.
- *Bayesian Network:* Similar to Naive Bayes classifier, the Bayesian Network (BN) is also a probability-based model used to predict the class label of the new instance of the dataset. It was used in cybersecurity to protect systems from unauthorized access. Cybersecurity is a challenging problem in Industry 4.0. Traditional security algorithms such as RSA, DES, and ECC can predict only

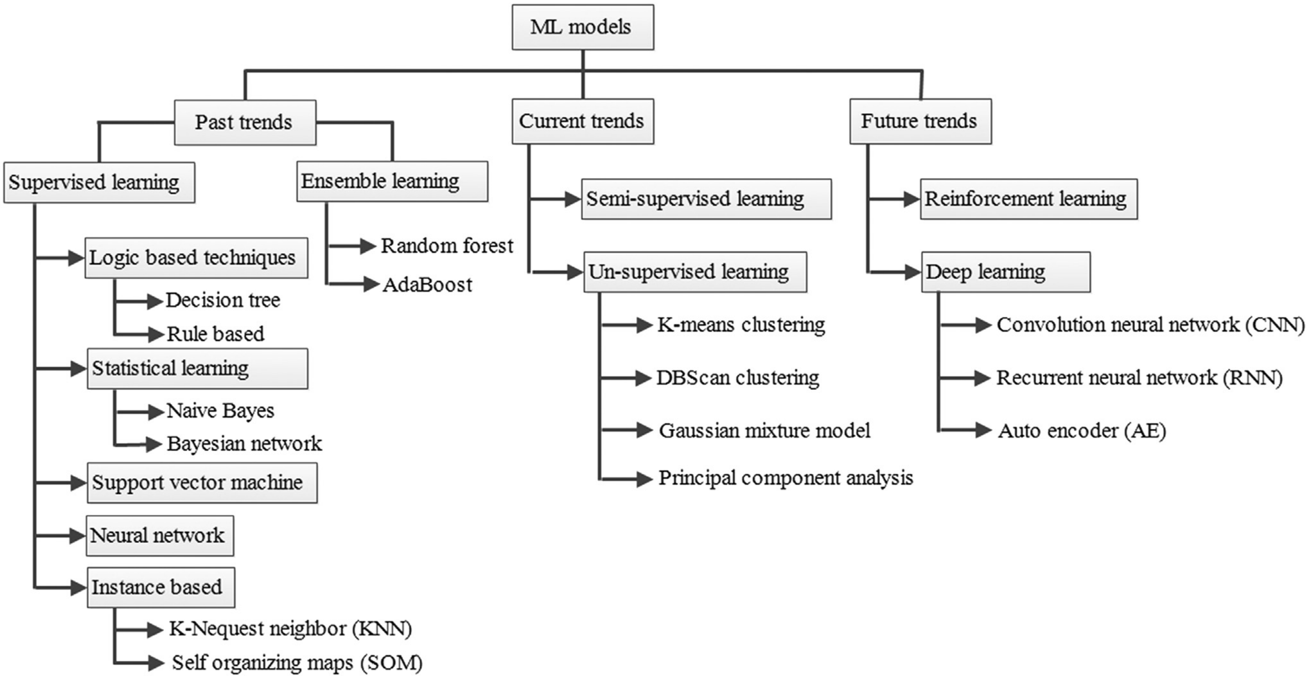


Fig. 11. Machine learning models for SDA.

known attacks and has no ability to learn from the traffic pattern to identify the new attacks. Zhang et al. [73] proposed a fuzzy-based probabilistic Bayesian Network for Industry 4.0 to provide dynamic cybersecurity. This system was designed with multi-domain knowledge of attacks and also designed a risk model from the historical information to predict the novel threats. Benefits of their system were: (i) effective risk assessment, (ii) effective at noisy instances, (iii) faster execution (4.90s with 490 nodes), and (iv) scalability.

BN was also used to detect false injected data at the packet level in wireless communication [74]. It's a physical layer technique to identify false data injection in a network packet, and the probability of false detection and false alarm decreases with an increase in the packet length. Sun et al. [75] used the BN to identify zero-day attack paths which were unknown to the public, i.e., new attack sequence. They proposed a ZePro probabilistic algorithm based on BN to identify a zero-day attack path. It establishes the instance graph, then it builds the BN using the instance graph to detect the intrusion activity. The authors concluded that the ZePro algorithm reconstructs the attack story, i.e., how the attack has occurred? Based on the storyline, it draws the instance graph and BN to highlight the zero-day attack path.

Support Vector Machine: It is a discriminative supervised ML approach with given labeled data to categorize the classes. It draws hyperplane in multidimensional space, which segregates data based on their classes. It can also be used to detect cyber attacks and network intruders based on the identification of their class labels. Ghanem et al. [76] used Support Vector Machine (SVM), which detects intrusion and reduces the number of false alarms. They have evaluated the performance of one-class and two-class SVM (both linear and non-linear forms) to identify the network intrusion. It has been observed that linear two-class SVM produces more accurate results over others. Its detection rate was almost 100% with linear two-class SVM classifier.

Authors in [77] proposed an efficient and effective network anomaly detection algorithm (SVM model), which guarantees a high level of network security. They used following network features to detect network traffic anomaly such as — source IP addresses count, source port count, destination IP addresses count, destination port numbers count, packet types, and packets of the same size. Feed these features into the SVM classifier, which then discriminate between the standard and attack data. They tested their algorithm with DARPA 1998 dataset and calculated the attack detection rate. The detection accuracy were 78.6%, 87.5%, and 82.4% for DoS, R2L, and U2R, respectively and for probing it was 87.3%.

Omrani et al. [78] used SVN classifier with ANN to detect network threats such as-TCP connection traffic. It is able to identify traffic at TCP connection as suspicious or normal, but too expensive. They evaluated the performance of their hybrid algorithm over the NSL-KDD DARPA dataset, which shows it was a promising model that increases the false detection rate and decreases the error rate.

Neural Network Model: Which is more intelligent — a human brain or computer? It motivates computer scientists to design a system that works like the human brain. Neural Network (NN) Model has a capability that behaves similarly to the human brain. It consists of thousands of neurons like "brain cells" in the human brain. There are different types of NNs such as-(i) feed-forward NN, (ii) RNN, (iii) modular NN, (iv) convolution NN, and (iv) Recursive NN. Network security threat detection with NN is the most adaptable application area nowadays. Demidova et al. [79] highlighted the use of NN in attack detection from network traffic. They classified the security attacks with the use of back-propagation NN (sigmoid activation function) on the net constructor interface. Then, Niu et al. [80] proposed a novel NN-based attack identification scheme capability to capture network abnormalities. This system generates a threat detection residual, which helps to determine the attack existence in the network traffic. The analysis was performed on MATLAB with parameters such as-sampling time (1 ms) and simulation time (200T). They found that the attack detection rate of their algorithm is faster than the traditional NN system, but it identifies only those network attacks, which causes delays and packet losses.

Instance-Based Models: It is also known as memory-based or lazy learning. It matches the closely resembled training instance with the new instance to detect the class label of new data. Various instance-based learning algorithms were K-Nearest Neighbor (KNN), Self Organization Map (SOM), and Learning Vector Quantization (LVQ). Discussion on some of the instance-based learning is explained in the subsequent paragraphs.

- *K-Nearest Neighbor:* It is a non-parametric based ML algorithm to solve classification problems, and its error classification rate is optimal with the increase in the value of K. K-Nearest Neighbor (KNN) classification model was used in network attack detection, where public key infrastructure and pseudonym certificates were used to secure confidentiality, integrity, but not adequate to handle Sybil attacks. Authors in [81] used the KNN-based ML classification model in threat identification and proposed a novel KNN-based Sybil attack detection in the vehicular system. It detected the driving pattern of the individual represented with eigenvalues and used to identify the Sybil attack based on the stored driving pattern instances. But, the problem with the KNN classifier is its runtime complexity. They have evaluated the performance on SUMO simulator with parameters — vehicle speed, road width, number of vehicles, the range of communication, and simulation time. The mean detection rate and accuracy of the algorithm is 80% and 100%, respectively. Moreover, Aung et al. [82] used the fusion of K-means with KNN to reduce the runtime complexity on the KNN classifier with higher accuracy. They used a hybrid model classifier for IDS to classify normal or attack data instances in KDD CUP 99 dataset. This hybrid model is also suitable in BD attack classification.
- *Self Organizing Maps*: Nowadays, data comprises text data, image data, voice data, video data, and map data. Except for the text data, all other data are high-dimensional data. Self Organizing Map (SOM) was well suited to visualize high-dimensional data in a highly competitive learning environment [83]. Applications of SOM were graph mining, social interaction, motion classifier, text-mining, network traffic analysis, and IDS. Dozono et al. [84] proposed a SOM-based model, i.e., *CGH-SOM* to analyze IP traffic using packet frequency. It was used to detect both DoS and DDoS attacks and improved the DoS attack detection accuracy. Moreover, similar work being carried by Almiani et al. [85] and developed a cluster-based SOM intelligent IDS, which works in two stages. In the first stage, the SOM network was building and in the second stage, apply agglomerative K-means clustering on neurons. It was used to address sensitivity and processing issues. This model was demonstrated on the NSL-KDD dataset and found sensitivity up to 96.66% and processing time up to 0.08 millisecond.

Table 8 shows the relative comparison of existing ML classification techniques concerning parameters such as proposed model, accuracy, detection rate, false alarm rate, the dataset used, pros, and cons.

Ensemble learning models. It is a method that combined multiple models to improve the performance and accuracy of the classifiers and also known as *committee learning* [88]. Authors in [89] used NN, decision tree (CART model), and regression to detect network intruders. The fusion of these classifiers improved the overall performance of IDS (99.94%) by voting of each classifier. They have tested the performance of an ensemble system with KDD CUP 1999 dataset. The common ensemble supervised learning methods were classified into random forest and AdaBoost models.

- *Random Forest Model (RF):* It was a statistical supervised ensemble learning technique based on grouping decision trees. It uses *Bagging* method to group the decision tree. It takes N-samples from the original sample with the use of the bootstrap method. Then, train multiple decision trees using these samples and predict the final result. Authors of [90] accessed this model with China National Internet Emergency Center (CNIEC) dataset. They observed the performance as 97% with RF is better as compared to Naive Bayes classifier (75%). Choi et al. [91] proposed the RF algorithm for IDS to optimize the number of decision trees used to make an accurate decision in detecting intruders and attacks. They evaluated their algorithm over KDD-CUP 99 dataset and found the accuracy as 99.97%, memory usage approx. 92% less and learning time 78.34s. It can also be used to detect distributed anomalies in wireless networks such as-(i) limited connectivity and (ii) real-time action response. Tsou et al. [92] designed a framework (OW-OCRF) using optimal weighted RF to overcome these aforementioned issues.
- *AdaBoost Model:* It is an ML algorithm to boost the weak ML algorithms, which slightly alters the training data and add the hypothesis to attain the highest accuracy [93]. Its computational runtime complexity is lower than that of the other ML algorithms. Authors in [94] used Naive Bayes as a weak classifier and boosted with AdaBoost ML algorithm. Their proposed algorithm extracts feature from the network traffic and feed it into the weak Naive Bayes classifier then apply AdaBoost with data tuning. It was used to detect Dos, U2R, R2L, and probe attacks and was experimented on KDD CUP 99 dataset and able to achieve 100% detection rate and approximately 0% false-positive rate.

Table 9 shows the relative comparison of existing ML ensemble learning techniques concerning parameters such as-proposed model, accuracy, detection rate, false alarm rate, the dataset used, pros, and cons.

7.3.2. Current-trends

This section highlights those ML techniques or models which are exhaustively being used to solve today's network security issues. Current trends in ML are classified into two categories: supervised and semi-supervised learning.

Unsupervised learning. It is a type of ML algorithm to draw inferences from unlabeled training data. The most adopted unsupervised learning algorithm is "*cluster analysis*". Other unsupervised learning algorithms are the Gaussian mixture algorithm, DBScan algorithm, and Hidden Markov Models. The various applications of unsupervised learning are fraud detection, data segregation, outliers detection, data compression, trend detection, and network security.

Clustering-Based Algorithms

- *K-Means Clustering:* It is a distance-based clustering algorithm which works in two steps-(i) calculate the cluster center and (ii) put the data in the nearest cluster center. Moreover, it helps to identify unknown patterns from unlabeled data. Sandhya et al. [98] proposed a conceptual framework based on the genetic K-means algorithm for intrusion detection (DoS, R2L, U2R, and probing). The genetic algorithm in clustering helps to identify optimal input features. They have used the NS2 simulator to implement the genetic algorithm to detect intruders in a wireless network using the AODV routing protocol. Parameters considered for the analysis were detection rate and false-positive rate. Their algorithm achieved a very high detection rate and low false-positive rate. This algorithm is competent in a dynamic environment and detects novel attacks without intrusion signatures. Later on, an improved version of the genetic K-means algorithm was presented by Sukumar et al. [99]. The number of clusters in an improved algorithm need not be fixed as in the traditional genetic K-means algorithm. It estimates the optimal value of K (number of clusters) using a fitness function. The evaluation was done on KDD-99 dataset and found that this modified K-means algorithm achieved accuracy with large datasets compared to the standard K-means algorithm, but its performance degrades with small datasets.

  Different variants of K-means algorithm has been presented by various authors such as min–max K-means [100], improved K-means [101], and modified K-means algorithm with chain clustering [102]. The agenda of the min–max K-means algorithm was to overcome the initial center sensitivity issues, cluster quality improvement, and intrusion detection. This min–max K-means algorithm minimizes the maximum intracluster variance. The

Table 8
Relative comparison of various ML classification techniques.

Author	Year	Model	Proposed model	Description	Attacks targetted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Hanmanthu et al. [66]	2015	DDL (Detection-Defense-Log) Model	DT classification model	Proposed a DT model to prevent SQL injection attack	SQL injection attack	✗	82%	✗	synthesis dataset	Prevention against SQL injection	Not discussed other possible attacks
Elekar et al. [69]	2015	Classification model, neural network model	Classification model for rule based	Presented five rule based classification algorithms such as JRip, Decision Table, PART, ZeroR, and OneR to detect network intrusion	DoS, Probe, U2R, R2L attacks	✗	90%	✗	KDD CUP1999 dataset	Better performance of PART classifier	Need to reduce false detection rate
Dozono et al. [84]	2015	SOM model	SOM model	Analyzed the network traffic of IP packets with CGH and SOM to detect malware attacks	Malware and DDoS attacks	✗	right(0.8)	✗	KDD CUP 99	Improved accuracy for Dosattack	Only discussed about DDoS attack
Jabbar et al. [65]	2016	Naive bayes model	Naive bayes model and alternating decision tree	Novel approach to classify intrusion attacks with ADT	Intrusion detection, network attacks	98.4 = DOS, 99.5= Probe, 96.7 = U2R and R2L	97.61 = DOS, 97.15 = Probe, 97.15 = U2R and R2L	3.3 = DOS, 5.5 = Probe, 2.38 = U2R and R2L	NSL KDD dataset	Classify various types of attacks effectively	Targeted only intrusion detection
Niu et al. [80]	2016	Adversary model	Non-linear neural network model	Their approach is capable to capture abnormality in the communication links.	cyberattacks and network attacks	✗	✗	✗	✗	Capture abnormality and efficient	only delays and packets losses
Lakshminarasimman et al. [67]	2017	J48, random forest decision tree algorithm	Decision tree classification model	Presented a DT classifier algorithm to detect anomaly and DDoS attacks	DDoS Attacks, hacking attacks, anomaly detection	✗	99.9415% = J48 DT algorithm, 96.9437% = Random Forest DT algorithm	✗	KDD-Cup99 dataset	Efficient anomaly detection	only detect DDoS attack
Gu et al. [81]	2017	Attack model	Attack model	Vehicles are classified based on the similarity in their driving patterns with using KNN algorithm	SybilAttack Detection	80% detection rate	80% accuracy	✗	✗	high detection ratio with a good performance in error control.	high runtime complexity

(*continued on next page*)

performance was evaluated over the NSL-KDD dataset with attack and standard input data and compare it with the traditional K-means algorithm [100]. The detection and false positive rate of min–max K-means algorithm was 81% and 9%, respectively, whereas, for traditional K-means, it is 75% and 14%, respectively. It works well with small datasets, and for large datasets, authors [102] modifies K-means algorithm with chain initialization with different window sizes based upon the size of the cluster. This modified algorithm is suitable to detect DDoS attacks in DARPA 98 dataset and its average performance was 98% detection rate, 99% accuracy, and 1.5% false-positive rate.

- *Gaussian Mixture Algorithm:* It was an unsupervised probabilistic learning model that discover class labels of the data with unknown parameters using finite Gaussian distribution. In this model, the process was used to identify abnormal or malicious

Table 8 (*continued*).

Author	Year	Model	Proposed model	Description	Attacks targetted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Lei [77]	2017	PSO-SVM model	PSO-SVM model	Presented the network anomaly detection algorithm based on SVM to detect different anomalies with high accuracy	DoS, Probe, U2R, and R2L	✗	Dos = 0.786, R2L = 0.875, U2R = 0.824, Probe = 0.873	✗	KDD-CUP'99 dataset	Higher accuracy	Only traffic anomaly detected
Omrani et al. [78]	2017	Markov model, probabilistic model, classification model	Probabilistic fusion model	Presented the fusion of ANN and SVM with NSL-KDD DARPA dataset to detect the attacks easily	DoS, Probe, U2R, and R2L and network attacks	79.71% detection TPR rate	High accuracy	better	NSL-KDD DARPA dataset	Combines both high and low level features and higher accuracy	Not able to detect each network connection attack.
Ghanem et al. [76]	2017	Classification model	Classification model	Presented the SVM to detect the cybersecurity attacks in network intrusion system	Cyber attacks, network attacks	100%	OSR reaches 81.67%, 99.25%, 93.51%	2.19%	network traffic datase, non-homogeneous dataset, Wi-fi dataset, probing dataset	reduce false alarm rate and better accuracy	False positive rate need to decrease more
Zhang et al. [73]	2018	Bow-tie model, Risk propagation model, Bayesian network model	FPBN model	Design a FPBN model to overcome the issue of limited historical data for the assessment of cybersecurity	✗	✗	$D_{min} = 1 \times 10^{-4}$.	not idea	✗	Removes the difficulty of limited historical data	No security and no attack targetted
Liu et al. [74]	2018	Bayesian network model	System model	Presented the Bayesian test for detecting the false data injection at the packet level in lossy one-way wireless relay	Byzantine attacks	PM(1-beta) + Pfbeta, missed detection for e <1/2	✗	false alarm for e < 1/2(E(U0) < E(U1))	✗	Detection of false data injection perfectly	No accuracy
Sun et al. [75]	2018	SODG model, Predecessor model, Propagation model, Local Observation Model	Object graph based and Bayesian network model	Proposed a model which identify zero-day attack paths in intrusion detection system	Zero-day attacks	✗	Better accuracy	81.38%	Patrol's dataset	successfully reveal the zero-day attack paths.	only reveal parts of the paths not capture the complete path
Aung et al. [82]	2018	Intrusion detection model	K-means and KNN model	Authors used k-means and KNN to reduce time complexity with great accuracy	zero-day or zero-hour attacks, DDoS, Probe, U2R, R2L	99%	99% accuracy	Low false alaram rate	KDDCup 99 dataset.	suitable in bigdata environment	Unknown threats not identified

(continued on next page)

packets in the network. This unusual or abnormal attack traffic was also called as *outliers*. Authors in [103] used a Gaussian Mixture (GM) algorithm to detect outliers from the input network traffic time series data. They have used the format of training data to identify the outliers from the test data (it should be in the same format). It detects outliers from both the historical and unseen data. It can also be applicable in BD to detect outliers. Some data sources systems were intelligent and need not be recognized as outliers, and such datasets harm the system performance. Bahrololum et al. [104] proposed a system with GM for anomaly detection, which learns the pattern of attack and normal data using Gaussian distribution. There were two phases in the GM model, such as-training and testing phase. In the training phase, the system extracts and learns the patterns of various attacks

Table 8 (*continued*).

Author	Year	Model	Proposed model	Description	Attacks targetted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Chen et al. [86]	2018	Intrusion detection model	KNN based intrusion detection model	Proposed a KNN and tree-seed algorithm (TSA) algorithm for feature selection to improve Classification efficiency of attacks	×	×	PSO = 83.65% GA = 78.78% TSA = 87.34%	×	KDD CUP 99 datasets	improve Classification efficiency and accuracy of intrusion detection	high runtime complexity
Yang et al. [72]	2018	Naive bayes model	Naive Bayes based artificial bee ant colony algorithm (ABNC)	Improve the performance of IDS using ABNC	×	×	91%	×	Sonar, Biodeg	IDS performance improved	Detection rate not calculated
Naseer et al. [87]	2018	Detection model, IDS model, ML model	Deep neural network IDS model	Presented anomaly detection model based on deep neural network	snmpget attack, DoS, Probe, U2R, R2L attacks	high detection rate	89%	×	NSLKDD dataset	Applicable in real-world anomaly detection systems.	Not efficient data representation
Xue et al. [70]	2018	Signature generation model, Bayes detection model	Signature generation model, Bayes detection model	Presented the PRSA algorithm with production rule inference mechanism to improve the detection rate and attack detection accurately	DoS, Probe, U2R, R2L attacks, network and cyber attacks, Apache-Knacker, BIND-TSIG and ATPhttpd	×	94%	×	KDD Cup 1999 test dataset	improved accuracy and reduces the false attack detection	No attack defense
Almiani et al. [85]	2018	Detection model, Classification model	Detection model	Presented a Intrusion detection system with clustered SOM	Network attacks	96.66%	83.46	×	NSL-KDD dataset, testing dataset	High computational efficiency	Weak sensitivity towards normal activities

and then design the classifier. In the testing phase, this modeled classifier reads the features extracted from the data input stream traffic and calculate either maximum probability or minimum deviation. This system has experimented with DARPA 98 dataset with chosen 310000 records and achieved the highest intrusion detection rate over the signature-based methods.

Apart from IDS, other attacks can also be detected using the GM model if implemented at the physical layer. Qiu et al. [105] proposed a physical layer authentication mechanism with a GM clustering approach to detect spoofing attackers. In this mechanism, the first step was to identify the parameters of the GM model using the expectation–maximization algorithm. Then, determine the deviations in input data packets to detect potential spoofing attackers. The maximum probability of detecting a spoofing attack was 0.98. Authors of [106] designed a hybrid decision tree-GM based system to detect misuse and anomaly, which can recognize attacks similar to the normal distribution. The detection rate, false-positive rate, and the accuracy of the hybrid model were better than the SVM classifier model.

- *DBScan Clustering:* It was a first density-based ML algorithm that divides the population into clusters based on the density of a region. It used *epsilon (eps-radius) and minpts (minimum number of neighbors)* as the parameters for data clustering. Various other density-based algorithms available were: Balanced Iterative Reducing and Clustering using hierarchies (BIRCH), Ordering Points to Identify the Clustering Structure (OPTICS), and Anomaly Detection with Fast Incremental Clustering (ADWICE). These density-based clustering algorithms were useful to identify network anomalies. Thang et al. [107] proposed a DBScan parameter identification mechanism which have used a different eps and minpts values known as DBScan-MP [107]. Comparison of DBScan-MP with original DBScan algorithm and also with other clustering algorithms was performed and found better in terms of intrusion detection rate and low false alarm rate in DBScan-MP with KDD CUP 99 dataset. Then, authors in [108] used DBScan to detect DDoS attacks in the wireless network, which identified the early phases of DDoS attack in unseen data. They have used two types of clustering: density-based (DBScan) and partitioning-based (K-means) clustering. Authors have experimented with their technique over CAIDA and DARPA 2000 datasets and achieved accuracy over 98%.
- *Principal Component Analysis:* It was a statistical mechanism used to convert the original data into compressed or reduced form without affecting data variances and dependencies. Moreover, it convert n correlated variables into d uncorrelated variables ($d < n$) called as principal components. The applications of PCA were dimensionality reduction and intrusion detection. But, the main problem with the network traffic is that it contains the most redundant and irrelevant information, which helps intruders to enter into the system. PCA resolved this problem, which transformed the large data into the reduced form [109]. It helps the

Table 9
Relative comparison of state-of-the-art ensemble learning techniques.

Author	Year	Proposed model	Description	Attacks targetted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Li et al. [95]	2010	Adaboost	Presented a statistical model based on Naive Bayes and AdaBoost classifiers to detect network intruders	DDoS, Probe, U2R, R2L attacks	100%	High accuracy	0%	KDD CUP 99.	Improved detection rate and low False Positive Rate	Need to execute with different datasets
Natesan et al. [94]	2012	Classification model Adaboost with J48 Base	Presented the Cascade classifier scheme based on Adaboostto enhance detection rate of rare network attack categories	Probe, U2R, R2L attacks	R2L = 59%, U2R = 45%, DoS and Probe = 85%	56.13%	–	KDD CUP 99	Detection rate of R2L and U2R increased individually	Accuracy reduced
Jin et al. [90]	2016	Classification model	Presented a Network Security assessment Model based on random forest classifier to improve accuracy	Cyberattacks	–	90%	–	–	More objective and accurate.	Only case study discussed
Ma et al. [92]	2017	preferential attachment (PA) model	Proposed a valid and robust solution to solve the social network unknown attack detection problem.	real-world de-anonymization, Sybil attack	–	–	0.5%	Kaggle dataset, real dataset	Good at noisy data,	High time complexity
Dong et al. [96]	2018	black-box, and white-box model	Presented an iterative models such as whitebox and blackbox models to boost adversial attacks	black box and white box attacks	–	–	–	Imagenet dataset	Robustness, Efficient	Not mentioned detection rate and accuracy values
Choi et al. [91]	2018	Memory efficienf RF model	Proposed an algorithm finds the stop point of the forest derivation using McNemar test	Probe, U2R, R2L attacks	–	100%	–	KDD-cup99 dataset.	Fast learning and training time	suitable for low power IDS
Tsou et al. [97]	2018	Random forest model	Presented a random forest based model to detect anomalies in WSN devices	Anomaly detection	–	–	–	multi-hop outdoor, Agriculture, and CRAC dataset	Allows resource constrained devices to work in collaboration and efficient	suitable for limited connections

system to extract only the most relevant information from the input data traffic. Hadri et al. [110] proposed an efficient IDS based on PCA and Fuzzy PCA, which eliminated the irrelevant information from the network input data. Then, they classified the input as an attack or normal with the KNN classifier. The experimental result proved that the Fuzzy PCA performs better than the PCA approach in attack detection because PCA has principal linear components which were more sensitive to noise and easily altered by noise.

To overcome an issue as mentioned above, authors used Robust Fuzzy PCA (RFPCA), which was used to reduce outliers affects. But, due to the large membership values, RFPCA still suffers from outliers. Then, the authors of [110] modified their approach using an RFPCA-based dimensionality reduction mechanism with the KNN classifier to make IDS more efficient and effective [111]. They conducted experiments on KDDCUP 99 and NSL-KDD dataset to evaluate its detection rate in comparison to RFPCA and PCA algorithms. Its performance was more promising with respect to RFPCA and PCA. Again they proposed Non-Linear based RFPCA (NFRPCA) to improve the performance of IDS [112]. NFRPCA detects all attacks and gives a very low false-positive rate. Almansob et al. [113] uses the Naive Bayes classifier instead of the KNN classifier to address the challenge to achieve a high detection rate, low false-positive rate, and high accuracy. They tested their algorithm over 41 feature (34-continuous and 7-discrete) KDD dataset. They were able to meet high rate values in the case of Naive Bayes classifier compared to the KNN classifier.

Table 10 shows the relative comparison of existing ML unsupervised learning techniques with reference to parameters such as-proposed model, accuracy, detection rate, false alarm rate, the dataset used, pros, and cons.

Semi-supervised learning. It adds the flavors of both supervised (labeled training data) and unsupervised (unlabeled training data) learning. The fusion of labeled and unlabeled training data considerably improved the accuracy of the learning algorithm. Supervised learning requires a labeled data, but it was hard to get labels on all the data instances in the training data because of the high cost and time-consuming labeling process. In situations where the cost and time are crucial, semi-supervised learning techniques can be preferred. It has many

Table 10
Relative comparison of state-of-the-art unsupervised learning techniques.

Author	Year	Proposed model	Description	Attacks targetted	Detection Rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Eslamnezhad et al. [100]	2014	Min–Max K-means	Introduced a min–max k-means clustering approach with NSL-KDD dataset in intrusion detection system	Dos, Pr, U2R, R2L attacks	DoS = 80.21%, U2R = 59.63%, R2L = 22.74%, Normal = 90.57%, Min–max = 81%, k-means = 75%	–	Min–max algo = 9%, k-means = 14%	NSL-KDD dataset	More efficient and high detection rate	Not much suitable for unknown attacks
Sandhya et al. [98]	2014	Genetic K-means	Proposed a framework to identify intruders in WSN using genetic k-means algorithm	Network attacks	Very high	–	Very low	Sensor data	Reliable, efficient, faster and more accurate	Not clearly specified percentage values
Pramana et al. [102]	2015	Modified K-means	Presented an approach of DoS detection with k-means algorithm with DARPA 98 dataset	DoS attacks	98%	99%,	1.50%	DARPA 98 dataset	Effectively detection of DoS attack	Only discussion about DoS attack
Alizadeh et al. [114]	2015	GMM model	Presented the use of GMM model for the verification and classification of network anomalies	zero-day attack	–	Flow based = 99.1%, Byte-based = 98.7%	–	Unibs-2009 dataset, public dataset	Use biometric verification methods for security	Only provide security against zero-days attack
Hadri et al. [110]	2016	PCA and Fuzzy PCA model	Proposed an IDS with PCA and FPCA algorithms to extract only useful features from dataset to improve accuracy	Pr, R2L, U2R DoS attack	DoS = 73.11%, U2R = 13.46%, R2L = 4.11%, Pr = 91.33%,	–	Worst false alaram rate	KDDcup99 dataset.	increase the detection rate and robustness	Poor false alarm rate value
Almansob et al. [63]	2017	Naive bayes and PCA model	Proposed a Naive bayes and PCA based model to solve issues of IDS	Pr, R2L, U2R DoS attack and others	98.82%	DoS = 87.47%, U2R = 95.76%, Pr = 97.12%, R2L = 98. 53%	0.52%	kdd99 dataset	reduce high dimensionality and provide feature extraction	Not tested against other datasets
Yin et al. [101]	2017	Improved K-means	Presented an improved K-means algorithm with information entropy and DD algorithm for Anomaly detection	back, smurf, teardrop, and neptune	98%	96.67%	0	KDD Cup99 and Iris dataset	efficient anomaly detection, provide higher accuracy rate.	Time consuming
Bitaab et al. [106]	2017	Gaussian Mixture Model (GMM)	Authors used decision tree for misuse detection and used Gaussian Mixture Model for anomaly detection	Hybrid intrusion detection	DT-GMM = 97.21% for dataset 1, DT-GMM = 96.72% for dataset 2	94.28%	8.59%	NSL-KDD dataset	Identify attacks similar to normal distribution	Cannot detect attacks on any other part of the network.
Hadri et al. [111]	2017	Robust Fuzzy PCA	To reduce outliers which effectively Identify network intruders robust Fuzzy PCA	Network attacks, Pr, R2L, U2R DoS	NSL-KDD dataset DoS = 73.2215%, U2R = 15.1123%, R2L = 4.7656%, Pr = 93.1128%	–	lowest	KDDCuP 99 dataset, NSL-KDD dataset	Improve the RF-PCA method.	Mode dimensions to be reduced for intruser detection

(continued on next page)

Table 10 (*continued*).

Author	Year	Proposed model	Description	Attacks targetted	Detection Rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Al-mamory et al. [108]	2017	DBSCAN model and detection model	Presented the DBSCAN Clustering Algorithm to detect network layer DDoSattacks.	DDoS attack	52.17%	98%	0.68%	DARPA 2000 dataset, CAIDA dataset	Provide early detection of DDoS attacks	Only discussed about DDoS attack
Reddy et al. [103]	2017	GMM model	Presented GMM model to Detect outliers in Network Traffic and BD scenario using probability density function	–	–	–	–	Big dataset, time series dataset, univariate, multivariate datasets	Outlier detection accurately	Threat event classification not done
Qiu et al. [105]	2018	GMM model	Presented the GMM based physical layer authentication approach to detect potential spoofer	Spoofing attack	>90%	High accuracy	–	Any transmitted message	achieve a very low Bayes risk	Only provide security against spoofing attack
Sukumar et al. [99]	2018	Modified Genetic K-means	Presented an improved genetic K-means (IGKM) algorithm for IDS	Dos, Pr, U2R, R2L attacks	–	k-means = 53.27%, IGKM = 72.91%	–	KDD-99 dataset	Suitable for large datasets, and provides higher accuracy	Not suitable for small datasets
Hadri et al. [112]	2018	Non-Linear Robust Fuzzy PCA	Presented the non-linear feature extraction mechanism called Nonlinear Fuzzy Robust PCA for IDS	Pr, R2L, U2R DoS attack	KDD-Cup99 dataset: DoS = 74.23%, U2R = 16.11%, R2L = 4.55%, Pr = 92.12%	–	Lowest	KDDCuP 99 dataset, NSL-KDD dataset	detection of all type of attacks.	Not well suited in real-time environment
Zhou et al. [115]	2019	MC-cluster tree	Designed a cluster tree to handle datasets dynamically with privacy preserving using edge computing	–	–	Higher average accuracy	–	Multimedia big datasets	Guarantee trust	Trade-off between privacy and accuracy

Pr: Probe.

applications in computer networks, and we discuss a few of them in the subsequent paragraphs.

Classification of unknown protocols having few labeled data can be done efficiently with a semi-supervised learning algorithm. Lin et al. [116] proposed a novel semi-supervised based learning approach for the situation, where a few labeled training data instances were available. This approach was suitable to detect unknown data samples generated by unknown network protocols. It works in three phases: such as-(i) sample label propagation, (ii) unknown protocol discovery, and (iii) protocol detailed classification. They have tested it against the three network traffic traces such as-WIDE-10, WIDE-12, and CND traces and found it significantly better compared to other ML algorithms.

Moreover, Divakaran et al. [117] proposed a self-learning classifier that learns a small number of data instances first, then reconstruct the classifier model with more number of instances. This self-learning classifier improves the accuracy of the classifier model to filter the network traffic as an attack or normal. It solves the fundamental issues of classification technique, i.e., dependence on labeled training data instances and not able to deal with non-static network input traffic. It achieved improvement in the accuracy of their learning model, and the zero-day attack was utterly avoided with this particular learning algorithm. Instead of analyzing only the network traffic, semi-supervised learning can also be used in IDS to achieve better performance and accuracy. Murthy et al. [118] analyzed the semi-supervised learning for IDS, which considerably saves both time and cost. They used three different classifiers to predict the output label. The implementation was done in the WEKA tool with NSL KDD and ISCX-botnet dataset and achieved a significant improvement in incorrect data classification accuracy.

7.3.3. Future-trends

This section discusses the futuristic ML algorithms and models such as-Reinforcement Learning (RL) and Deep Learning (DL). The detailed description of RL and DL algorithms are explained in the next paragraphs.

Reinforcement learning. It is a class of ML algorithm which maximize rewards in a specific situation. It is different from both supervised and unsupervised learning algorithms. RL does not require any training data; it learns from experience. Model designing in supervised and unsupervised learning is quite easy because of the presence of a training dataset. Decisions are dependent on the previous output in RL, whereas independent in case of supervised learning. The example would be a chess game is an example of RL and object recognition is an example of supervised learning. RL can be positive or negative, where positive RL increases behavior strength and frequency if it causes some event and negative RL means to strengthen the behavior due to condition stopped.

Various applications of RL are in robotics, data processing, creating self-instruction training systems, and in network security to detect network intrusions. Nowadays, RL becomes more popular in the area of network security applications. Randrianasolo et al. [120] proposed

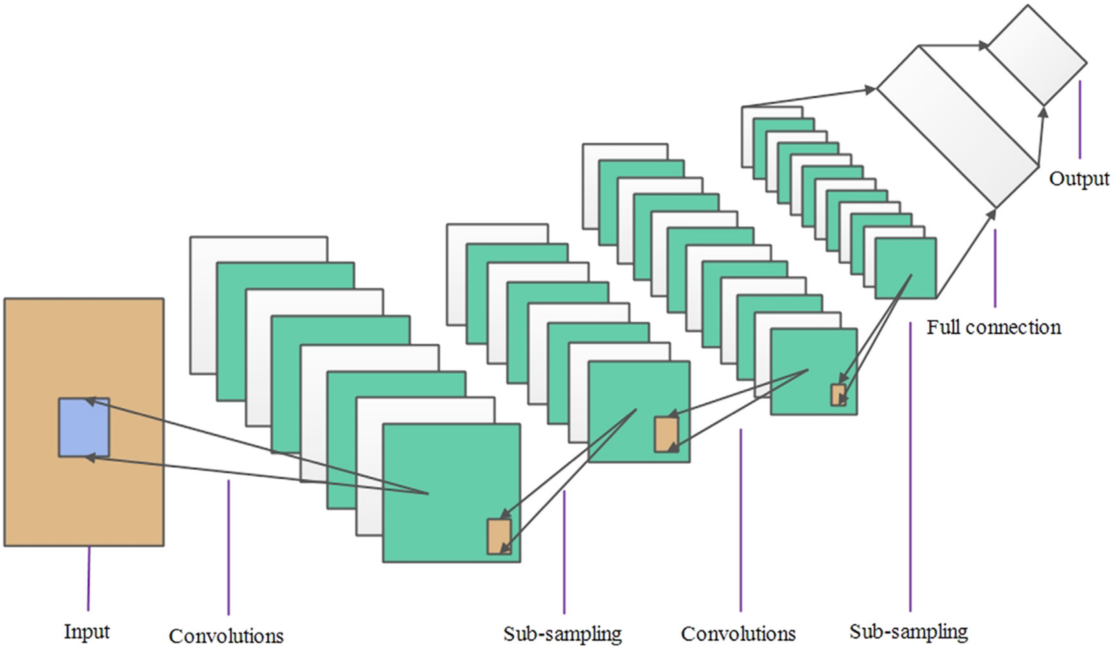


Fig. 12. Convolution neural network structure [119].

a Q-learning RL algorithm to secure software that learns itself for temporary repair of the network. In this algorithm, an environment was represented with a tree, where a node represents a state and branch represents an action. It calculates the Q-values of the links and compares it with the fixed threshold value, if it falls below the set threshold value, then software security design needs improvement. This approach was used to block malicious activities.

Further, Yousefi et al. [121] analyzed the Q-learning based attack graph model for Multi-Host Multi-Stage Vulnerability Analysis (MulVAL) to generate an attack graph from the network topology. It simplified and got the transition graph, which models the environment. Then, apply a Q-learning algorithm to identify possible attack paths that help the attacker to breach the network security. This technique helps network administrators to take possible actions to mitigate unknown security threats. It is also used to make IDS intelligent, efficient, and accurate in an anonymous or novel environment. Authors in [122] proposed a Q-learning based decision system to control the processes in an undetermined environment. It identifies the optimal strategy to block intruders with an immediate optimal response. They have evaluated the effectiveness of the system with Root Mean Square Error and found its better compared to the [121] approach.

Deep learning. It is an improved version on ML techniques that learns and extracts features using multiple interlinked layers. It extracts notable features from data only, whereas ML accepts features from an expert/user. DL is successfully implemented in the areas of image processing, natural language processing, object recognition, voice synthesis, and speech recognition. Various DL models available are Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM), and Auto-Encoders (AE). It takes more time to train the network [123,124], as compared to the traditional ML models.

- *Convolution Neural Network:* It was first developed to solve an image recognition problem and achieved better results. Now, people started using CNN to address network traffic and security-related issues, i.e., classification of network traffic data. Zhou et al. [119] proposed a modified traditional CNN algorithm to improve the traffic classification. It uses min–max normalization to evaluate the incoming traffic data then map it into a grayscale image, which feeds as an input to the convolution layer. This algorithm improved the accuracy and reduce classification time compared to the traditional CNN and useful to extract in-depth features from the input traffic dataset. DL is widely used in object or image recognition and not been used to detect physical security threats. Authors [125] used CNN with transfer learning to reduce the training time and increase the threat detection accuracy by a physical security system. They evaluated GoogLeNet, AlexNet, and Quadratic SVM with transfer learning for physical security systems and found GoogLeNet (100% accuracy) and AlexNet (few frames misclassified) achieves better accuracy. It has been observed that their algorithm was well suited for some specific images in physical security assessment, but accuracy can be increased with an increase in training dataset size.
CNN can also be used for network intrusion detection. Vinayakumar et al. [126] analyzed the Transmission Control Protocol (TCP)/Internet Protocol (IP) with DL supervised learning algorithms such as CNN, RNN, LSTM, and Grated Recurrent Unit (GRU). They experimented their analysis over KDD Cup 99 dataset with 1000 epochs to evaluate the efficiency, but this approach was not suitable to identify new threats and analyze live network traffic. Later, the efficiency and accuracy of the IDS with CNN can be increased with the usage of a grayscale encoding scheme [127]. In this, numeric features were encoded with 10-digit binary representation, whereas sub-category features were with a single digit. Using this scheme, a numeric feature occupied 1.25 pixels, whereas other occupied 0.125 pixels. Its performance was analyzed with TensorFlow, UNSW-NB15, and IDS2017 datasets over the Inception V3 model with parameters epoch, learning rate, and batch size. They observed that the CNN-based anomaly detection method achieved better accuracy in comparison to RF classifier in certain scenarios and applications.
In this paragraph, we discuss the working of CNN, which aims to learn appropriate interrelated features from the input data to identify possible threats. It has a basic nature of *weight sharing and pooling* (*max/min/average*). The function of weight sharing is to learn the same pattern throughout the image, whereas the function of the pooling layer is to reduce the dimensions of input data and also gives positional and translational invariance. Each CNN layer is composed of convolution kernels that generate feature maps [128]. After a few convolutions and pooling layers, one or few fully connected layers are used for the classification

of input data. Each convolution layers use an activation function such as ReLU, Leaky ReLU, and Sigmoid. The diagrammatic representation of CNN along with its convolution and pooling layers is as shown in Fig. 12.

- *Recurrent Neural Network & Long-Short Term Memory:* It is a type of NN in which following output is to be calculated from previous output, whereas in traditional NN all inputs and outputs are independent (have no connection with each other). Applications of Recurrent Neural Network (RNN) are in the field of natural language processing, video, and speech recognition. It has limited memory which stores the immediately previous output which is used to predict the next output word or sequence. But, there are some situations, where the only immediate output is not sufficient to predict the next output. For example, "predicting the next number in sequence", then the single previous number is not enough for the prediction because numbers are in series. For accurate prediction, we need to store a few previous outputs. RNN is not suitable for the aforementioned example but, LSTM is well suited for this particular situation which stores a few outputs instead on only previous output. Architectural flow of both RNN and LSTM is as shown in Figs. 14 and 13.

  Let $S_{T\text{-}1}$ be the activated output of the hidden layer at time *t*, W_{HL} be the weight vector of the hidden layer, W_{OL} be the weight vector of the output layer, W_{to_next} be the weight used for next hidden layer, Inp_T be the input word vector, Out_T be the output word vector at different timestamps. Eq. (1) and (2) represents the flow of RNN with activation (ActFun) and Softmax function.

$$S_{\mathrm{T}} = ActFun(W_{\mathrm{HL}} * Inp_{\mathrm{T}} + W_{to\_next} * S_{\mathrm{T}-1}) \tag{1}$$

$$Out_{\mathrm{T}} = Softmax(W_{\mathrm{OL}} * S_{\mathrm{T}}) \tag{2}$$

  These networks are used to analyze and predict the network intruders and malware by matching the previously-stored malicious activities. Many authors worldwide are working on it. Yin et al. [5] proposed an RNN-based Network Intrusion Detection System (RNN-IDS) which preprocess the data before training. They compared its performance with ANN, RF, and SVM and claimed that the RNN-IDS delivered a higher accuracy and superior classification. Fu et al. [129] modified the existing RNN-IDS model with LSTM-RNN to enhance the accuracy of the classification system and proposed an LSTM-RNN-based system to achieve highly accurate attack detection, data processing, feature abstraction, and training. Data preprocessing helps to provide high-quality data for processing. Authors have evaluated their model on the NSL-KDD dataset and calculate its detection rate (98.85%), false alarm rate (8.75%), and accuracy (97.52%). It's performance parameters have better values with LSTM-RNN compared to GRNN, Probabilistic Neural Network (PNN), KNN, and SVM. Moreover, authors of [130] tested the LSTM-RNN CSIC 2010 HTTP dataset, which has 36000 standard data and 25000 attack data. They have considered parameters such as-learning rate (0.01), number of hidden layers (6), epochs (100), and batch size (50) and achieved the accuracy as 99.97%.

- *Auto Encoders:* Various DL models use available datasets, but it contains many irrelevant data dimensions. The Auto Encoders (AE) concept of DL solved this issue by reducing the data into lower dimensions without losing its semantics. It is also called a dimensionality reduction technique, which improves the accuracy of the system and also reduces its computational complexity. It enhances the security of the system against the network intruders [131]. The network vulnerabilities are increasing with a rapid rate, i.e., both known and unknown vulnerabilities, which cannot be easily identified by any IDS. We need to reduce the data dimensions to get appropriate data for further processing. It can be possible with Deep AE (DAE). Farahnakian et al. [132] proposed a DAE-based IDS which trained the system layer-wise to reduce over-fitting and maximize optimization. The performance was evaluated and achieved the detection accuracy as 94.71% with KDD Cup 99 dataset.

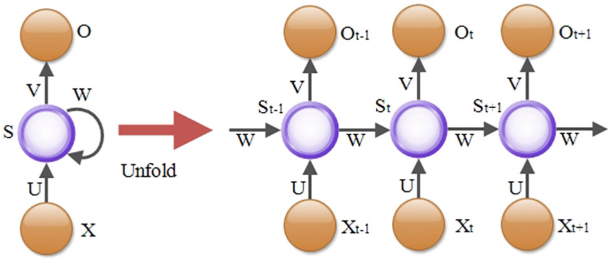


Fig. 13. Recurrent Neural Network working model [5].

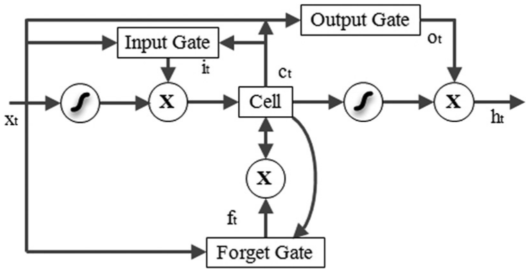


Fig. 14. Long-Short Term Memory cell architecture [129].

Table 11 shows the relative comparison of existing ML deep learning techniques with reference to parameters such as-proposed model, attacks targeted, accuracy, detection rate, false alarm rate, the dataset used, pros, and cons.

7.4. Biometric-based SDA

Maintaining the security of data during the information exchange mechanism over the network is a prime concern. It can be achieved with the use of traditional cryptographic algorithms such as-Rivest Shamir Adleman (RSA), Secure Hash Algorithm (SHA), Elliptic Curve Cryptography (ECC), Advanced Encryption Standard (AES) and data encryption standard (DES), and the smart cards. But, the problem with cryptographic algorithms is the *management of private keys* whereas, the problem with the smart card is that it might be lost or stolen [138]. These issues can be rectified with the use of measurable physical or behavioral characteristics of the user for authentication and verification purposes, which makes the system more secure without the need to remember complex passwords and private keys is known as *biometrics* [139]. The various biometric methods available are fingerprint identification, face recognition, retina scan, iris scan, voice analysis, and hand geometry. These methods are more secured in the present era to maintain privacy and security.

The use of biometrics in various sensitive fields such as — healthcare records [140], military applications, voting system, and aviation to protect the system against unauthorized access. He et al. [138] proposed a biometric-based scheme for system authentication with the usage of ECC. It is a three-factor authentication scheme (password, smart card, and biometrics) that protects against various attacks such as-replay attack, man-in-the-middle attack, and modification attack. It reduces both the computation and communication cost. This approach mainly focused on performance factors instead of security. Authors in [141] designed a secure cloud to manage healthcare records, which ensures security to medical data access with biometric-based authentication mechanisms. In this scheme, the authors used an ML back-propagation NN to train biometric signature samples. It prioritizes

Table 11
Comparison of state-of-the-art Deep Learning Techniques.

Author	Year	Proposed model	Description	Attacks targeted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Vinayakumar et al. [126]	2017	Analysis of CNN model	Analyzed the performance of CNN for network IDS by using TCP/IP protocol for predefined time-series with KDDCup 99 dataset	Dos, Probe, U2R, R2L attacks	–	>0.98% of CNN layer-3 LSTM	>0.0 of CNN layer 2 RNN	KDDCup 99 dataset	Improved performance	High computation cost
Stubbs et al. [125]	2017	Transfer Learning over CNN	Used transfer learning over CNN for physical security assessment to detect alarm in the physical security system	Dos, Probe, U2R, R2L attacks	–	100%	0.8% = GoogleNet, 0% = AlexNet	ImageNet, AlexNet, and Quadratic SVM dataset	Increased stability and physical security assessment	Not suited for large datasets
Zhou et al. [119]	2017	Min-Max CNN	Presented the traffic classification algorithm based on improved CNN	–	–	99.3%	–	Traffic dataset	Improved accuracy and reduced classification time	No attack detection
Yin et al. [5]	2017	RNN-IDS model	Presented the deep- learning approach to create a RNN-IDS model for intrusion detection	Dos, Probe, U2R, R2L attacks	Dos = 83.49%, Probe = 83.40%, U2R = 11.50%, R2L = 24.69%	KDDTrain = 99.53%, KDDtest = 81.29%,	Dos = 2.06%, Probe = 2.16%, U2R = .07%,	NSL-KDD dataset.	Improved accuracy of IDS and recognize intrusion type.	High training time
Yeo et al. [133]	2018	CNN model	Introduced a malware detection system with CNN for higher accuracy, precision, and recall values	Malware attacks	–	>85%	–	Keras dataset	Robust and accurate malware identification	Detection and false alarm rate not considered
Kim et al. [127]	2018	GoogleNet, CNN model	Presented an approach of DoS detection with k-means algorithm over DARPA 98 dataset	DDoS, ransomware, DoS, Web, infiltration, botnet, and port scan attacks	–	>89%	–	NSL-KDD, UNSW-NB15, IDS2017 dataset	Anomaly detection performance enhanced	Not always works better
Wang et al. [134]	2018	CNN model	Presented a CNN based representation learning mechanism for identification of malicious traffic	Ddos and network attacks	–	Multi-tag classifier = 96.75%, binary classifier = 97.3%, F1 score = 98.5%.	–	UNSW-NB15, KDD99 dataset, Tensorflow	High accuracy in malicious traffic detection	Works only with small datasets
Teoh et al. [135]	2018	DL models	Applied RNN and J48 DL algorithms in cybersecurity space for threat identification	malicious and cyber-attacks	–	RNN = 0.964%	–	Raw dataset	Achieved good accuracy	No false rate and detection rate
Fu et al. [129]	2018	LSTM RNN model	Presented an effective network attack detection method with RNN to achieve data processing, feature abstraction, and training	Network attack, SQL-based attacks, intrusion attack, DOS, U2R, R2L, Probe.	98.85%	97.52%	8.75%	NSL-KDD dataset, NSL-KDD — a benchmark dataset	Faster, achieve data processing, and high detection rate	Can be better with deep RNN
Meng et al. [136]	2018	PCA and LSTM RNN model	Analyzed the network threat with PCA and LSTM-RNN	SQL-based, intrusion, DOS, U2R, R2L, Probe, and malicious attacks	>90%	98.85%	4.86%	NSL-KDD dataset	Preserves attack features of input traffic data	Detection rate can be increased with more dimensionality reduction

(*continued on next page*)

Table 11 (*continued*).

Author	Year	Proposed model	Description	Attacks targeted	Detection rate	Accuracy	False alaram rate	Dataset used	Pros	Cons
Althubiti et al. [130]	2018	LSTM-RNN model	Applied LSTM-RNN for intrusion detection to predict known and unknown attacks	Cyber-attacks, Malicious attack, Probe, R2L, U2R, and DoS attack	98.95%	97.54%	9.98%	CSIC 2010 HTTP dataset	Efficient IDS binary classifier	Need optimization in attack detection
Vartouni et al. [137]	2018	Deep AE model	Proposed a deep AE model for anomaly detection and web attacks	Web attacks	88.34%	88.32%	–	CSIC 2010 dataset	deep model show better performance.	Need input as data stream for performance evaluation
Abolhasanzadeh [131]	2018	Non-linear probabilistic model	Presented a non-linear dimensionality reduction mechanism to detect intruders with AE features	Cyber and network attacks	–	better accuracy	–	KDDCuP 99 and NSL-KDD dataset	Efficient	High time complexity
Farahnakian et al. [132]	2018	Auto-Encoder (AE) IDS model	Presented a Deep Auto-encoder approach for IDS with KDD-CUP'99 dataset	Probe, R2L, U2R DoS attack, Zero-days, network attacks	Binary classification = 95.65%, Multi-classification = 94.53%	Binary classification = 96.53%, Multi-classification = 94.71%	Binary classification = 0.35%, Multi-classification = 0.42%	KDD-CUP 99 dataset	Avoid overfitting and local optima, high accuracy and detection rate	Sparsity constraints are not applied

the processing of key records and suitable for large data applications also. But, not adequate to handle all possible attacks on BD and their mitigation strategies. Moreover, authors in [142] presented a secure and lightweight biometric scheme using ECC to protects against the various malicious attacks such as-replay attack, impersonation attack, server spoofing attack, privileged insider attack, user anonymity, modification attack, and man-in-the-middle attack. It also provides low communication and computational cost. They have evaluated its performance with Samsung Galaxy S6 smartphone with data rate 5.76 MB and different ECC key sizes such as-160, 192, 224, and 256 bits.

Biometric-based security schemes are frequently used in Healthcare 4.0. Advancements in healthcare technologies, i.e., implantable devices (ID), improves human life expectancy [143]. These IDs are used to function various organs of the human body in an artificially perfect way [144]. Doctors have the right to monitor the IDs of authorized patients through a communication link (i.e., Bluetooth) and diagnose the patients accordingly. These IDs are accessed through a network, so it is more susceptible to attacks. For example, any unauthorized access to IDs can malfunction its working and also data leakage. A secure authentication mechanism (with ECC) for ID was proposed by Wazid et al. [144] to mitigate security and privacy issues in ID communication. They verified it against the known attacks with the practical demonstration on NS2 simulator and secured with physical characteristics authentication such as biometrics.

Table 12 shows the relative comparison of existing biometric-based secure EHR techniques with reference to parameters such as-environment, authentication, attacks targeted, pros, and cons.

7.5. Unmanned aerial vehicle-assisted SDA

It is often called as the drone, and its usage has been increased in the last few years. It is used in areas for surveillance, where human life risk involvement is quite high [145]. It executes costly and dangerous missions effectively and also attracted the focus of various industries and academics in multiple applications such as-surveillance, package delivery, and traffic monitoring [146]. Data in Unmanned Aerial Vehicle (UAV), like other connected devices, are susceptible to various cyber-attacks such as-confidentiality, integrity, and availability (CIA). Categorization of numerous cyber-attacks are shown in Fig. 3.

Authors in [22] analyzed the cyberattacks on the remotely collected sensor data and proposed a secure solution. They have given the safeguard solutions to ensure availability, confidentiality, and integrity. They used a firewall to ensure availability, public key infrastructure, and certificates to ensure the integrity and symmetric encryption for data confidentiality. A variety of attacks are targeted by different authors working in the same area. But, their proposed systems are able to identify only those attacks which are there in the database. It does not restrict the novel or unknown attacks because of not specified attack definition. ML-based architecture helps to predict both known as well as unknown attacks for attack-free real-time video streaming and analysis.

7.6. Infrastructure-based SDA

In online information sharing, security remains paramount. Normally, anomalies can be detected through pattern matching with traffic monitoring [147]. The data volume is increasing day-by-day exponentially. So, traffic analysis needs more accurate as well as real-time monitoring and computation power. A traditional monitoring mechanism has some issues, such as (i) scalability, (ii) data security, and (ii) identification of new threats. To resolve the aforementioned issues, more secured infrastructures are required, such as tool, framework, or model. Yang et al. [147] proposed a highly scalable framework to detect threat anomalies with ML data models such as-statistical data models, i.e., linear regression, which resolved the traffic data monitoring issue in BDA. BD can be used in every field of organization such as-financial and transportation. There may be a chance of frauds in BD financial transactions which may cause data corruption, data disclosure, and violate data integrity which attracted cybercriminals to perform some malicious activities.

Authors in [148] presented a sentinel tool to secure direct debit payments, which can analyze frauds against the direct debit payments. The experimental results show the improvement in data processing with calculated values of the performance parameters such as (i) throughput of 676 tuples/sec, (ii) latency of 28 ms, (iii) number of queries per second are 6000, and (iv) response time of 3.5 ms. Vegh [149] proposed a framework to access control and perform secure data analytics. This framework analyzes, prevent, detect, and mitigate cyberattacks on BD

Table 12
Relative comparison of biometric-based secure EHR approached.

Author	Year	Objective	Environment	Authentication	Attacks targeted	Pros	Cons
He et al. [138]	2014	Presented a biometric and ECC-based authentication scheme for distributed multiserver environment	Multi-server distributed environment	ECC-based biometric authentication	Replay, MIM, privileged insider, and impersonation attacks	Suited for multiserver network environments	No scalability
Shakil et al. [141]	2017	Proposed a biometric-based authentication and data management scheme for healthcare with digital signature	Cloud distributed environment	Digital signature-based biometric authentication	Replay, MIM, privileged insider, impersonation, and DoS attacks	Highly scalable, robust, and flexible	Hard to identify intruder
Mohammedi et al. [142]	2018	Designed a biometric-based remote patient authentication and monitoring approach	Remote healthcare monitoring environment	Biometric reader with biometric template	Replay, impersonation, DoS, Spoofing, and modification attacks	Light weight and well suited for mobile platforms	Possibility of MD stolen attack
Wazid et al. [144]	2018	Proposed a secure remote authentication framework for Implantable medical devices	Implantable environment	Fuzzy extractor biometric authentication scheme	Replay and man-in-the-middle attack	Increased throughput, low communication cost, and reduce delay	Network reliability not considered

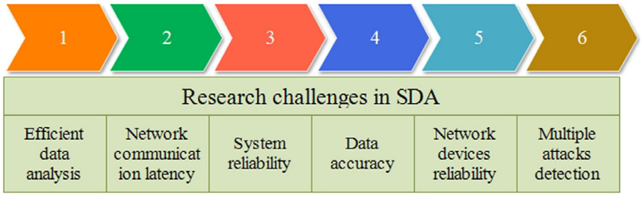

1
2
3
4
5
6
Research challenges in SDA
Efficient data analysis
Network communicat ion latency
System reliability
Data accuracy
Network devices reliability
Multiple attacks detection

Fig. 15. Research challenges in ML-based SDA.

systems. But, it does not focus on traditional security methods such as-encryption and steganography. An authentication factor is used as a security mechanism in this framework and would be facial recognition, fingerprint, and voice recognition. Other security mechanisms used as token passing and answer a security question and a user can choose multiple authentication factors at a time.

7.7. Web-based SDA

The development of the Internet was for communication, i.e., send and receive messages. It is used to connect families, friends, and officials from remote locations. But, some people use the Internet to steal credentials of users such as-password and bank account details. So, providing security to users personal data is one of the major concerns in digital communication. As the data is increasing, the risk associated with the data is also increasing. Nowadays, small databases are not adequate to store such a huge volume of data. For that, we need BD (capable of storing huge amount of heterogeneous data).

Various security mechanisms are available to secure user credentials such as cryptography and hashing techniques. Ahmed et al. [150] designed a hash-based method to protect user credentials on web browser against identity theft. This method requires no change on the server and minimum change at the client-side. It works only in Google Chrome, not in other browsers such as Firefox, opera, and safari.

> This sections described the taxonomy which shows the various ML-models for SDA. It highlighted the ML-trend wise (past, current, and future trends) ML models used for data analytics security. Past trend includes supervised and ensemble learning, current trend is having un-supervised and semi-supervised learning, whereas future trends includes deep learning reinforcement learning. It also highlighted the bifurcation based on other features also.

8. Research challenges in SDA

The discussion of SDA for various application areas in context to data analysis security. Rapid advancements in information technology encourage the process and system automation in almost all the areas such as-smart home, smart city, manufacturing, logistics, and healthcare. This leads to rapid and exponential data growth, i.e., data generated in every millisecond. The devices such as-sensor devices generates a huge amount of real-time data and send it to the cloud for further processing and analysis. Existing systems use fog and mobile edge computing to reduce latency and increases data security. But still, there is an issue of security with the analysis of data being captured from heterogeneous sources such as-sensor devices, Internet, machines, smart grid, and enterprises. Based on the exhaustive literature survey, we identified some issues and challenges in ML-based SDA implementation that need to be addressed, as shown in Fig. 15.

- *Efficient data analysis:* It is a major concern in SDA system design if we look for efficiency, then we need to compromise with the data security. Before sending data to the cloud, it should be encrypted for secure processing. Before the processing and analysis of data, it needs to be decrypted. The encryption–decryption process is time-consuming, which reduces the efficiency of data analysis. It has been observed from the literature that both the data efficiency and security are inversely proportional to each other and can be achieved with the implementation of ML algorithm (*supervised, unsupervised, NN, CNN, and AE*) which helps to detect network intruders and anomalies from network traffic without any cryptographic methods.
- *Network communication latency:* To reduce delay in network communication is also a challenge in designing the SDA system. More the delay (existing network communication delay), the more the chances of network attacks which may compromise the data security. 5G-enabled Tactile Internet (TI) is a viable solution for the aforementioned issue which ensures <1ms latency and *99.999%* reliability, and availability [151]. Rapid transfer of data makes the data secure and reliable for analysis.
- *System Reliability:* Reliability of data is also a prime concern to achieve correct processing and analysis results. Input traffic data is categorized into either normal or attack data. Correct classification of input traffic data is important otherwise, analysis results may vary. Traditional network security algorithms are not a viable solution to classify input traffic data as attack or normal

because they mainly focused on data security, but not on network traffic filtering. It can be done with the usage of ML models (DT, ensemble learning, CNN, AE, and RNN), which learns itself from experience and classify the data, which makes processing results reliable.

- *Data accuracy:* As we know, the input traffic data can be either suspicious or normal data. Accuracy of data analysis and processing result depends upon the type of input data. To achieve higher accuracy with traditional security algorithms is a challenge in SDA system design. ML algorithms are a viable solution to classify the input traffic as normal or suspicious. Moreover, these algorithms help the SDA system to produce comparatively accurate processing results.
- *Network devices reliability:* Nowadays, the data is exponentially growing, which requires large storage. The personal systems are not capable of processing and storing such a tremendous amount of data (petabyte or zettabyte). So, people use cloud or fog distributed storage to store, analyze, and process the data. The storage is at remote location, and access is through the network communication channel. Thus, the trustworthiness and reliability of network devices in a distributed network is a challenge to ensure Quality of Service (QoS). An effective trust-based ML model needs to be explored for SDA to resolve the aforementioned issues.
- *Multiple attacks detection:* To date, many researchers have used ML algorithms to identify the network traffic abnormalities or some attacks such as-DoS, DDoS, probe, spoofing, and back-door attacks. Still, its a challenge for the researchers to mitigate all possible attacks against the systems.

> This section presented the various research challenges that can be faced using ML-based models for SDA.

9. Conclusion

In this paper, we provide insights to the readers about the use of ML models to secure data for further analytics. The survey is divided into four parts. The first part discussed about the background of SDA and security attacks on ML models in detail. Next, the architecture components of SDA along with both physical and logical interfaces of the traditional and ML-based SDA architecture are analyzed. The second part discussed about the two phases; ML models and SDA. In the first phase, ML model trends, threat model, and their countermeasures are discussed. In the second phase, SDA infrastructure, network traffic, SDA processing, policies for SDA, and UAV assisted SDA are discussed. The comparative analysis of the existing ML models used for SDA is performed on the basis of types of attacks, targeted area, and the parameters to mitigate these attacks. Finally, the fourth part discussed about open issues and challenges of using ML models in SDA.

In future, the more real-time attacks on the ML models would be explored in detailed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by Visvesvaraya Ph.D. Scheme for Electronics and IT by Department of Electronics and Information Technology (DeiTY), Ministry of Communications and Information Technology, Government of India.

References

[1] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, R.M. Parizi, K.-K.R. Choo, Fog data analytics: A taxonomy and process model, J. Netw. Comput. Appl. 128 (2019) 90–104, http://dx.doi.org/10.1016/j.jnca.2018.12.013.

[2] E.M.C. Digital Universe, The digital universe of opportunities: Rich data and the increasing value of the Internet of Things, 2014, https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

[3] P. Sangani, Global data to increase 10× by 2025: Data age 2025, 2017, https://economictimes.indiatimes.com/tech/internet/global-data-to-increase-10x-by-2025-data-age-2025/articleshow/58004862.cms?from=mdr.

[4] A. Jindal, A.K. Marnerides, A. Scott, D. Hutchison, Identifying security challenges in renewable energy systems: A wind turbine case study, in: Proceedings of the Tenth ACM International Conference on Future Energy Systems, e-Energy ’19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 370–372, http://dx.doi.org/10.1145/3307772.3330154.

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

[6] M. Mayhew, M. Atighetchi, A. Adler, R. Greenstadt, Use of machine learning in big data analytics for insider threat detection, in: MILCOM 2015 - 2015 IEEE Military Communications Conference, 2015, pp. 915–922, http://dx.doi.org/10.1109/MILCOM.2015.7357562.

[7] J. Hu, A.V. Vasilakos, Energy big data analytics and security: Challenges and opportunities, IEEE Trans. Smart Grid 7 (5) (2016) 2423–2436, http://dx.doi.org/10.1109/TSG.2016.2563461.

[8] N. Chaudhari, S. Srivastava, Big data security issues and challenges, in: 2016 International Conference on Computing, Communication and Automation, ICCCA, 2016, pp. 60–64, http://dx.doi.org/10.1109/CCAA.2016.7813690.

[9] J. Gardiner, S. Nagaraja, On the security of machine learning in malware c&c detection: A survey, ACM Comput. Surv. 49 (3) (2016) 59:1–59:39, http://dx.doi.org/10.1145/3003816.

[10] N. Singh, D.P. Singh, B. Pant, A comprehensive study of big data machine learning approaches and challenges, in: 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS, 2017, pp. 80–85, http://dx.doi.org/10.1109/ICNGCIS.2017.14.

[11] O. Yavanoglu, M. Aydos, A review on cyber security datasets for machine learning algorithms, in: 2017 IEEE International Conference on Big Data, Big Data, 2017, pp. 2186–2193, http://dx.doi.org/10.1109/BigData.2017.8258167.

[12] F. Jiang, Y. Fu, B.B. Gupta, F. Lou, S. Rho, F. Meng, Z. Tian, Deep learning based multi-channel intelligent attack detection for data security, IEEE Trans. Sustain. Comput. (2018) 1, http://dx.doi.org/10.1109/TSUSC.2018.2793284.

[13] P. Mishra, V. Varadharajan, U. Tupakula, E.S. Pilli, A detailed investigation and analysis of using machine learning techniques for intrusion detection, IEEE Commun. Surv. Tutor. 21 (1) (2019) 686–728, http://dx.doi.org/10.1109/COMST.2018.2847722.

[14] M. Husak, J. Komarkova, E. Bou-Harb, P. Celeda, Survey of attack projection, prediction, and forecasting in cyber security, IEEE Commun. Surv. Tutor. 21 (1) (2019) 640–660, http://dx.doi.org/10.1109/COMST.2018.2871866.

[15] R.A.A. Habeeb, F. Nasaruddin, A. Gani, I.A.T. Hashem, E. Ahmed, M. Imran, Real-time big data processing for anomaly detection: A survey, Int. J. Inf. Manage. (2018) http://dx.doi.org/10.1016/j.ijinfomgt.2018.08.006.

[16] M.S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, A.P. Sheth, Machine learning for internet of things data analysis: a survey, Digit. Commun. Netw. 4 (3) (2018) 161–175, http://dx.doi.org/10.1016/j.dcan.2017.10.002.

[17] D. Liu, Z. Yan, W. Ding, M. Atiquzzaman, A survey on secure data analytics in edge computing, IEEE Internet Things J. 6 (3) (2019) 4946–4967, http://dx.doi.org/10.1109/JIOT.2019.2897619.

[18] S. Sobati Moghadam, A. Fayoumi, Toward securing cloud-based data analytics: A discussion on current solutions and open issues, IEEE Access 7 (2019) 45632–45650, http://dx.doi.org/10.1109/ACCESS.2019.2908761.

[19] S. Tanwar, Verification and validation techniques for streaming big data analytics in Internet of Things environment, IET Netw. (2018).

[20] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, G. Loukas, A taxonomy and survey of attacks against machine learning, Comp. Sci. Rev. 34 (2019) 100199, http://dx.doi.org/10.1016/j.cosrev.2019.100199.

[21] J. Singh, Real time big data analytic: Security concern and challenges with machine learning algorithm, in: 2014 Conference on IT in Business, Industry and Government, CSIBIG, 2014, pp. 1–4, http://dx.doi.org/10.1109/CSIBIG.2014.7056985.

[22] H. Benkraouda, E. Barka, K. Shuaib, Cyber attacks on the data communication of drones monitoring critical infrastructure, Acad. Ind. Res. Collab. Cent. J. (2018) 83–93.

[23] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – A systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15, http://dx.doi.org/10.1016/j.infsof.2008.09.009, Special Section - Most Cited Articles in 2002 and Regular Research Papers.

[24] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, 2007.

[25] P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, J. Syst. Softw. 80 (4) (2007) 571–583, http://dx.doi.org/10.1016/j.jss.2006.07.009, Software Performance.

[26] A. Gupta, R.K. Jha, Security threats of wireless networks: A survey, in: International Conference on Computing, Communication Automation, 2015, pp. 389–395, http://dx.doi.org/10.1109/CCAA.2015.7148407.

[27] S. Tanwar, J. Vora, S. Tyagi, N. Kumar, M.S. Obaidat, A systematic review on security issues in vehicular ad hoc network, Secur. Priv. 1 (5) (2018) e39, http://dx.doi.org/10.1002/spy2.39.

[28] S. Banerjee, V. Odelu, A.K. Das, S. Chattopadhyay, N. Kumar, Y. Park, S. Tanwar, Design of an anonymity-preserving group formation based authentication protocol in global mobility networks, IEEE Access 6 (2018) 20673–20693, http://dx.doi.org/10.1109/ACCESS.2018.2827027.

[29] A. Srivastava, S.K. Singh, S. Tanwar, S. Tyagi, Suitability of big data analytics in Indian banking sector to increase revenue and profitability, in: 2017 3rd International Conference on Advances in Computing,Communication Automation, ICACCA (Fall), 2017, pp. 1–6, http://dx.doi.org/10.1109/ICACCAF.2017.8344732.

[30] A. Jindal, A. Schaeffer-Filho, A. Marnerides, P. Smith, A. Mauthe, L. Granville, Tackling energy theft in smart grids through data-driven analysis, in: IEEE ICNC 2020, IEEE, 2019.

[31] A. Saleem, A. Khan, S.U.R. Malik, H. Pervaiz, H. Malik, M. Alam, A. Jindal, FESDA: Fog-enabled secure data aggregation in smart grid IoT network, IEEE Internet Things J. (2019) 1, http://dx.doi.org/10.1109/JIOT.2019.2957314.

[32] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M.S. Obaidat, J.J.P.C. Rodrigues, Fog computing for smart grid systems in the 5G environment: Challenges and solutions, IEEE Wirel. Commun. 26 (3) (2019) 47–53, http://dx.doi.org/10.1109/MWC.2019.1800356.

[33] V.R. More, B.K. Patil, R.A. Auti, Secure extraction of association rules in horizontally distributed database using improved unifi, in: 2017 International Conference on Big Data Analytics and Computational Intelligence, ICBDAC, 2017, pp. 205–210, http://dx.doi.org/10.1109/ICBDACI.2017.8070835.

[34] J. Vora, P. Italiya, S. Tanwar, S. Tyagi, N. Kumar, M.S. Obaidat, K. Hsiao, Ensuring privacy and security in e-health records, in: 2018 International Conference on Computer, Information and Telecommunication Systems, CITS, 2018, pp. 1–5, http://dx.doi.org/10.1109/CITS.2018.8440164.

[35] V.R.R. Atukuri, R.S.R. Prasad, A novel approach: Reliable and secure data storage and retrieval in a cloud, in: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS, 2017, pp. 1296–1300, http://dx.doi.org/10.1109/ICECDS.2017.8389653.

[36] S. Bothe, A. Cuzzocrea, P. Karras, A. Vlachou, Skyline query processing over encrypted data: An attribute-order-preserving-free approach, in: Proceedings of the First International Workshop on Privacy and Secuirty of Big Data, PSBD '14, ACM, New York, NY, USA, 2014, pp. 37–43, http://dx.doi.org/10.1145/2663715.2669613.

[37] A. Cuzzocrea, A reference architecture for supporting secure big data analytics over cloud-enabled relational databases, in: 2016 IEEE 40th Annual Computer Software and Applications Conference, COMPSAC, vol. 2, 2016, pp. 356–358, http://dx.doi.org/10.1109/COMPSAC.2016.224.

[38] D. Puthal, S. Nepal, R. Ranjan, J. Chen, A secure big data stream analytics framework for disaster management on the cloud, in: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems, HPCC/SmartCity/DSS, 2016, pp. 1218–1225, http://dx.doi.org/10.1109/HPCC-SmartCity-DSS.2016.0170.

[39] S. Ojha, V. Rajput, Aes and md5 based secure authentication in cloud computing, in: 2017 International Conference on I-SMAC, IoT in Social, Mobile, Analytics and Cloud, I-SMAC, 2017, pp. 856–860, http://dx.doi.org/10.1109/I-SMAC.2017.8058300.

[40] G. Murali, R.S. Prasad, Secured cloud authentication using quantum cryptography, in: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS, 2017, pp. 3753–3756, http://dx.doi.org/10.1109/ICECDS.2017.8390166.

[41] V.R.R. Atukuri, R.S.R. Prasad, A novel approach: Reliable and secure data storage and retrieval in a cloud, in: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS, 2017, pp. 1296–1300, http://dx.doi.org/10.1109/ICECDS.2017.8389653.

[42] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, Fog computing for healthcare 4.0 environment: Opportunities and challenges, Comput. Electr. Eng. 72 (2018) 1–13, http://dx.doi.org/10.1016/j.compeleceng.2018.08.015.

[43] R.K. Barik, H. Dubey, A.B. Samaddar, R.D. Gupta, P.K. Ray, FogGIS: Fog computing for geospatial big data analytics, in: 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering, UPCON, 2016, pp. 613–618, http://dx.doi.org/10.1109/UPCON.2016.7894725.

[44] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M. Maasberg, K.-K.R. Choo, Multimedia big data computing and Internet of Things applications: A taxonomy and process model, J. Netw. Comput. Appl. 124 (2018) 169–195, http://dx.doi.org/10.1016/j.jnca.2018.09.014.

[45] R. Gupta, S. Tanwar, S. Tyagi, N. Kumar, Tactile internet and its applications in 5G era: A comprehensive review, Int. J. Commun. Syst. 32 (14) (2019) e3981, http://dx.doi.org/10.1002/dac.3981, e3981 dac.3981.

[46] S. Tanwar, J. Vora, S. Kaneriya, S. Tyagi, Fog-based enhanced safety management system for miners, in: 2017 3rd International Conference on Advances in Computing,Communication Automation, ICACCA (Fall), 2017, pp. 1–6, http://dx.doi.org/10.1109/ICACCAF.2017.8344726.

[47] J. Vora, S. Tanwar, S. Tyagi, N. Kumar, J.J.P.C. Rodrigues, FAAL: Fog computing-based patient monitoring system for ambient assisted living, in: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, Healthcom, 2017, pp. 1–6, http://dx.doi.org/10.1109/HealthCom.2017.8210825.

[48] F. Mehdipour, B. Javadi, A. Mahanti, FOG-Engine: Towards big data analytics in the fog, in: 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DASC/PiCom/DataCom/CyberSciTech, 2016, pp. 640–646, http://dx.doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2016.116.

[49] L. Hernandez, H. Cao, M. Wachowicz, Implementing an edge-fog-cloud architecture for stream data management, in: 2017 IEEE Fog World Congress, FWC, 2017, pp. 1–6, http://dx.doi.org/10.1109/FWC.2017.8368538.

[50] C. Dsouza, G. Ahn, M. Taguinod, Policy-driven security management for fog computing: Preliminary framework and a case study, in: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014, 2014, pp. 16–23, http://dx.doi.org/10.1109/IRI.2014.7051866.

[51] T.D. Dang, D. Hoang, A data protection model for fog computing, in: 2017 Second International Conference on Fog and Mobile Edge Computing, FMEC, 2017, pp. 32–38, http://dx.doi.org/10.1109/FMEC.2017.7946404.

[52] D. Liu, Z. Yan, W. Ding, M. Atiquzzaman, A survey on secure data analytics in edge computing, IEEE Internet Things J. (2019) 1, http://dx.doi.org/10.1109/JIOT.2019.2897619.

[53] H. Cui, X. Yi, S. Nepal, Achieving scalable access control over encrypted data for edge computing networks, IEEE Access 6 (2018) 30049–30059, http://dx.doi.org/10.1109/ACCESS.2018.2844373.

[54] P. Zhou, K. Wang, J. Xu, D. Wu, Differentially-private and trustworthy online social multimedia big data retrieval in edge computing, IEEE Trans. Multimed. (2018) 1, http://dx.doi.org/10.1109/TMM.2018.2885509.

[55] S. Garg, A. Singh, K. Kaur, G.S. Aujla, S. Batra, N. Kumar, M.S. Obaidat, Edge computing-based security framework for big data analytics in vanets, IEEE Netw. 33 (2) (2019) 72–81, http://dx.doi.org/10.1109/MNET.2019.1800239.

[56] A. Alabdulatif, I. Khalil, X. Yi, Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption, J. Parallel Distrib. Comput. 137 (2020) 192–204, http://dx.doi.org/10.1016/j.jpdc.2019.10.008.

[57] S. Saxena, DISHA: Data ownership, security, consent for health data, 2018, https://novojuris.com/2018/08/16/disha-data-ownership-security-consent-for-health-data/?utm_source=Mondaq&utm_medium=syndication&utm_campaign=inter-article-link.

[58] N. Legal, India: Data localisation: India's policy framework, 2018, http://www.mondaq.com/india/x/739546/Data+Protection+Privacy/Data+Localisation+Indias+Policy+Framework.

[59] S. Chandra, S. Ray, R.T. Goswami, Big Data Security in Healthcare: Survey on Frameworks and Algorithms, in: 2017 IEEE 7th International Advance Computing Conference, IACC, 2017, pp. 89–94.

[60] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, M. Saadi, Big data security and privacy in healthcare: A review, Procedia Comput. Sci. 113 (2017) 73–80, http://dx.doi.org/10.1016/j.procs.2017.08.292, The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)/The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017)/Affiliated Workshops.

[61] C. Bachelet, M. Kettani, K. Idrissi, Data protection laws of the world, Morocco, 2019, www.dlapiperdataprotection.com.

[62] Trunomil, The EU General Data Protection Regulation (GDPR) is the most important change in data privacy regulation in 20 years, 2018, https://eugdpr.org.

[63] Agri-Analyticsl, Agri-analytics: Harvesting agri-business insight, 2018, https://agrianalytics.com.au/what-is-it/.

[64] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P.K. Singh, W. Hong, Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward, IEEE Access 8 (2020) 474–488, http://dx.doi.org/10.1109/ACCESS.2019.2961372.

[65] M.A. Jabbar, S. Samreen, Intelligent network intrusion detection using alternating decision trees, in: 2016 International Conference on Circuits, Controls, Communications and Computing, I4C, 2016, pp. 1–6, http://dx.doi.org/10.1109/CIMCA.2016.8053265.

[66] B. Hanmanthu, B.R. Ram, P. Niranjan, Sql injection attack prevention based on decision tree classification, in: 2015 IEEE 9th International Conference on Intelligent Systems and Control, ISCO, 2015, pp. 1–5, http://dx.doi.org/10.1109/ISCO.2015.7282227.

[67] S. Lakshminarasimman, S. Ruswin, K. Sundarakantham, Detecting DDoS attacks using decision tree algorithm, in: 2017 Fourth International Conference on Signal Processing, Communication and Networking, ICSCN, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICSCN.2017.8085703.

[68] T. Komviriyavut, P. Sangkatsanee, N. Wattanapongsakorn, C. Charnsripinyo, Network intrusion detection and classification with decision tree and rule based approaches, in: 2009 9th International Symposium on Communications and Information Technology, 2009, pp. 1046–1050, http://dx.doi.org/10.1109/ISCIT.2009.5341005.

[69] K. Elekar, M.M. Waghmare, A. Priyadarshi, Use of rule base data mining algorithm for intrusion detection, in: 2015 International Conference on Pervasive Computing, ICPC, 2015, pp. 1–5, http://dx.doi.org/10.1109/PERVASIVE.2015.7087051.

[70] M. Xue, W. Yu, An attack signatures generation sequence alignment algorithm based on production rules, in: 2018 10th International Conference on Communication Software and Networks, ICCSN, 2018, pp. 270–274, http://dx.doi.org/10.1109/ICCSN.2018.8488280.

[71] D.H. Deshmukh, T. Ghorpade, P. Padiya, Intrusion detection system by improved preprocessing methods and Naïve Bayes classifier using NSL-KDD 99 Dataset, in: 2014 International Conference on Electronics and Communication Systems, ICECS, 2014, pp. 1–7, http://dx.doi.org/10.1109/ECS.2014.6892542.

[72] J. Yang, Z. Ye, L. Yan, W. Gu, R. Wang, Modified naive Bayes algorithm for network intrusion detection based on artificial bee colony algorithm, in: 2018 IEEE 4th International Symposium on Wireless Systems Within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS, 2018, pp. 35–40, http://dx.doi.org/10.1109/IDAACS-SWS.2018.8525758.

[73] Q. Zhang, C. Zhou, Y. Tian, N. Xiong, Y. Qin, B. Hu, A fuzzy probability Bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems, IEEE Trans. Ind. Inf. 14 (6) (2018) 2497–2506, http://dx.doi.org/10.1109/TII.2017.2768998.

[74] X. Liu, Y. Guan, S.W. Kim, Bayesian test for detecting false data injection in wireless relay networks, IEEE Commun. Lett. 22 (2) (2018) 380–383, http://dx.doi.org/10.1109/LCOMM.2017.2771274.

[75] X. Sun, J. Dai, P. Liu, A. Singhal, J. Yen, Using Bayesian networks for probabilistic identification of zero-day attack paths, IEEE Trans. Inf. Forensics Secur. 13 (10) (2018) 2506–2521, http://dx.doi.org/10.1109/TIFS.2018.2821095.

[76] K. Ghanem, F.J. Aparicio-Navarro, K.G. Kyriakopoulos, S. Lambotharan, J.A. Chambers, Support vector machine for network intrusion and cyber-attack detection, in: 2017 Sensor Signal Processing for Defence Conference, SSPD, 2017, pp. 1–5, http://dx.doi.org/10.1109/SSPD.2017.8233268.

[77] Y. Lei, Network anomaly traffic detection algorithm based on SVM, in: 2017 International Conference on Robots Intelligent System, ICRIS, 2017, pp. 217–220, http://dx.doi.org/10.1109/ICRIS.2017.61.

[78] T. Omrani, A. Dallali, B.C. Rhaimi, J. Fattahi, Fusion of ANN and SVM classifiers for network attack detection, in: 2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering, STA, 2017, pp. 374–377, http://dx.doi.org/10.1109/STA.2017.8314974.

[79] Y. Demidova, M. Ternovoy, Neural network approach of attack's detection in the network traffic, in: 2007 9th International Conference - the Experience of Designing and Applications of CAD Systems in Microelectronics, 2007, pp. 128–129, http://dx.doi.org/10.1109/CADSM.2007.4297500.

[80] H. Niu, S. Jagannathan, Neural network-based attack detection in nonlinear networked control systems, in: 2016 International Joint Conference on Neural Networks, IJCNN, 2016, pp. 4249–4254, http://dx.doi.org/10.1109/IJCNN.2016.7727754.

[81] P. Gu, R. Khatoun, Y. Begriche, A. Serhrouchni, k-nearest neighbours classification based sybil attack detection in vehicular networks, in: 2017 Third International Conference on Mobile and Secure Services, MobiSecServ, 2017, pp. 1–6, http://dx.doi.org/10.1109/MOBISECSERV.2017.7886565.

[82] Y.Y. Aung, M. Myat Min, Hybrid intrusion detection system using k-means and k-nearest neighbors algorithms, in: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science, ICIS, 2018, pp. 34–38, http://dx.doi.org/10.1109/ICIS.2018.8466537.

[83] V.K. Pachghare, P. Kulkarni, D.M. Nikam, Intrusion detection system using self organizing maps, in: 2009 International Conference on Intelligent Agent Multi-Agent Systems, 2009, pp. 1–5, http://dx.doi.org/10.1109/IAMA.2009.5228074.

[84] H. Dozono, N. Okada, The analysis of traffic of IP packets using CGH self organizing maps, in: 2015 International Conference on Computational Science and Computational Intelligence, CSCI, 2015, pp. 215–219, http://dx.doi.org/10.1109/CSCI.2015.55.

[85] M. Almi'ani, A.A. Ghazleh, A. Al-Rahayfeh, A. Razaque, Intelligent intrusion detection system using clustered self organized map, in: 2018 Fifth International Conference on Software Defined Systems, SDS, 2018, pp. 138–144, http://dx.doi.org/10.1109/SDS.2018.8370435.

[86] F. Chen, Z. Ye, C. Wang, L. Yan, R. Wang, A feature selection approach for network intrusion detection based on tree-seed algorithm and k-nearest neighbor, in: 2018 IEEE 4th International Symposium on Wireless Systems Within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems, IDAACS-SWS, 2018, pp. 68–72, http://dx.doi.org/10.1109/IDAACS-SWS.2018.8525522.

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

[88] P. Sornsuwit, S. Jaiyen, Intrusion detection model based on ensemble learning for u2r and r2l attacks, in: 2015 7th International Conference on Information Technology and Electrical Engineering, ICITEE, 2015, pp. 354–359, http://dx.doi.org/10.1109/ICITEED.2015.7408971.

[89] R. Kumar Singh Gautam, E.A. Doegar, An ensemble approach for intrusion detection system using machine learning algorithms, in: 2018 8th International Conference on Cloud Computing, Data Science Engineering, Confluence, 2018, pp. 14–15, http://dx.doi.org/10.1109/CONFLUENCE.2018.8442693.

[90] Y. Jin, Y. Shen, G. Zhang, The model of network security situation assessment based on random forest, in: 2016 7th IEEE International Conference on Software Engineering and Service Science, ICSESS, 2016, pp. 977–980, http://dx.doi.org/10.1109/ICSESS.2016.7883229.

[91] S. Choi, D. Ko, S. Hwang, Y. Choi, Memory-efficient random forest generation method for network intrusion detection, in: 2018 Tenth International Conference on Ubiquitous and Future Networks, ICUFN, 2018, pp. 305–307, http://dx.doi.org/10.1109/ICUFN.2018.8436590.

[92] J. Ma, Y. Qiao, G. Hu, Y. Huang, A.K. Sangaiah, C. Zhang, Y. Wang, R. Zhang, De-anonymizing social networks with random forest classifier, IEEE Access 6 (2018) 10139–10150, http://dx.doi.org/10.1109/ACCESS.2017.2756904.

[93] S. Kaneriya, S. Tanwar, S. Buddhadev, J.P. Verma, S. Tyagi, N. Kumar, S. Misra, A range-based approach for long-term forecast of weather using probabilistic Markov model, in: 2018 IEEE International Conference on Communications Workshops, ICC Workshops, 2018, pp. 1–6, http://dx.doi.org/10.1109/ICCW.2018.8403541.

[94] P. Natesan, P. Rajesh, Cascaded classifier approach based on adaboost to increase detection rate of rare network attack categories, in: 2012 International Conference on Recent Trends in Information Technology, 2012, pp. 417–422, http://dx.doi.org/10.1109/ICRTIT.2012.6206789.

[95] W. Li, Q. Li, Using naive Bayes with AdaBoost to enhance network anomaly intrusion detection, in: 2010 Third International Conference on Intelligent Networks and Intelligent Systems, 2010, pp. 486–489, http://dx.doi.org/10.1109/ICINIS.2010.133.

[96] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9185–9193, http://dx.doi.org/10.1109/CVPR.2018.00957.

[97] Y. Tsou, H. Chu, C. Li, S. Yang, Robust distributed anomaly detection using optimal weighted one-class random forests, in: 2018 IEEE International Conference on Data Mining, ICDM, 2018, pp. 1272–1277, http://dx.doi.org/10.1109/ICDM.2018.00171.

[98] S. G, A. Julian, Intrusion detection in wireless sensor network using genetic k-means algorithm, in: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 1791–1794, http://dx.doi.org/10.1109/ICACCCT.2014.7019418.

[99] J.V. Anand Sukumar, I. Pranav, M. Neetish, J. Narayanan, Network intrusion detection using improved genetic k-means algorithm, in: 2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI, 2018, pp. 2441–2446, http://dx.doi.org/10.1109/ICACCI.2018.8554710.

[100] M. Eslamnezhad, A.Y. Varjani, Intrusion detection based on minmax k-means clustering, in: 7'th International Symposium on Telecommunications, IST'2014, 2014, pp. 804–808, http://dx.doi.org/10.1109/ISTEL.2014.7000814.

[101] C. Yin, S. Zhang, J. Wang, J. Kim, An improved k-means using in anomaly detection, in: 2015 First International Conference on Computational Intelligence Theory, Systems and Applications, CCITSA, 2015, pp. 129–132, http://dx.doi.org/10.1109/CCITSA.2015.11.

[102] M.I.W. Pramana, Y. Purwanto, F.Y. Suratman, Ddos detection using modified k-means clustering with chain initialization over landmark window, in: 2015 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC, 2015, pp. 7–11, http://dx.doi.org/10.1109/ICCEREC.2015.7337056.

[103] A. Reddy, M. Ordway-West, M. Lee, M. Dugan, J. Whitney, R. Kahana, B. Ford, J. Muedsam, A. Henslee, M. Rao, Using Gaussian mixture models to detect outliers in seasonal univariate network traffic, in: 2017 IEEE Security and Privacy Workshops, SPW, 2017, pp. 229–234, http://dx.doi.org/10.1109/SPW.2017.9.

[104] M. Bahrololum, M. Khaleghi, Anomaly intrusion detection system using Gaussian mixture model, in: 2008 Third International Conference on Convergence and Hybrid Information Technology, vol. 1, 2008, pp. 1162–1167, http://dx.doi.org/10.1109/ICCIT.2008.17.

[105] X. Qiu, T. Jiang, S. Wu, M. Hayes, Physical layer authentication enhancement using a Gaussian mixture model, IEEE Access 6 (2018) 53583–53592, http://dx.doi.org/10.1109/ACCESS.2018.2871514.

[106] M. Bitaab, S. Hashemi, Hybrid intrusion detection: Combining decision tree and Gaussian mixture model, in: 2017 14th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology, ISCISC, 2017, pp. 8–12, http://dx.doi.org/10.1109/ISCISC.2017.8488375.

[107] T.M. Thang, J. Kim, The anomaly detection by using dbscan clustering with multiple parameters, in: 2011 International Conference on Information Science and Applications, 2011, pp. 1–5, http://dx.doi.org/10.1109/ICISA.2011.5772437.

[108] S.O. Al-mamory, Z.M. Algelal, A modified dbscan clustering algorithm for proactive detection of DDoS attacks, in: 2017 Annual Conference on New Trends in Information Communications Technology Applications, NTICT, 2017, pp. 304–309, http://dx.doi.org/10.1109/NTICT.2017.7976107.

[109] S. Tanwar, T. Ramani, S. Tyagi, Dimensionality reduction using PCA and SVD in big data: A comparative case study, in: Z. Patel, S. Gupta (Eds.), Future Internet Technologies and Trends, Springer International Publishing, Cham, 2018, pp. 116–125.

[110] A. Hadri, K. Chougdali, R. Touahni, Intrusion detection system using PCA and fuzzy PCA techniques, in: 2016 International Conference on Advanced Communication Systems and Information Security, ACOSIS, 2016, pp. 1–7, http://dx.doi.org/10.1109/ACOSIS.2016.7843930.

[111] A. Hadri, K. Chougdali, R. Touahni, Identifying intrusions in computer networks using robust fuzzy PCA, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications, AICCSA, 2017, pp. 1261–1268, http://dx.doi.org/10.1109/AICCSA.2017.78.

[112] A. Hadri, K. Chougdali, R. Touahni, A network intrusion detection based on improved nonlinear fuzzy robust PCA, in: 2018 IEEE 5th International Congress on Information Science and Technology, CiSt, 2018, pp. 636–641, http://dx.doi.org/10.1109/CIST.2018.8596643.

[113] S.M. Almansob, S.S. Lomte, Addressing challenges for intrusion detection system using naive Bayes and PCA algorithm, in: 2017 2nd International Conference for Convergence in Technology, I2CT, 2017, pp. 565–568, http://dx.doi.org/10.1109/I2CT.2017.8226193.

[114] H. Alizadeh, A. Khoshrou, A. Zúquete, Traffic classification and verification using unsupervised learning of Gaussian mixture models, in: 2015 IEEE International Workshop on Measurements Networking, M N, 2015, pp. 1–6, http://dx.doi.org/10.1109/IWMN.2015.7322980.

[115] P. Zhou, K. Wang, J. Xu, D. Wu, Differentially-private and trustworthy online social multimedia big data retrieval in edge computing, IEEE Trans. Multimed. 21 (3) (2019) 539–554, http://dx.doi.org/10.1109/TMM.2018.2885509.

[116] R. Lin, O. Li, Q. Li, Y. Liu, Unknown network protocol classification method based on semi-supervised learning, in: 2015 IEEE International Conference on Computer and Communications, ICCC, 2015, pp. 300–308, http://dx.doi.org/10.1109/CompComm.2015.7387586.

[117] D.M. Divakaran, L. Su, Y.S. Liau, V.L.L. Thing, SLIC: Self-learning intelligent classifier for network traffic, Comput. Netw. 91 (2015) 283–297, http://dx.doi.org/10.1016/j.comnet.2015.08.021.

[118] A. Jaiswal, A.S. Manjunatha, B.R. Madhu, P.C. Murthy, Predicting unlabeled traffic for intrusion detection using semi-supervised machine learning, in: 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques, ICEECCOT, 2016, pp. 218–222, http://dx.doi.org/10.1109/ICEECCOT.2016.7955218.

[119] H. Zhou, Y. Wang, X. Lei, Y. Liu, A method of improved CNN traffic classification, in: 2017 13th International Conference on Computational Intelligence and Security, CIS, 2017, pp. 177–181, http://dx.doi.org/10.1109/CIS.2017.00046.

[120] A.S. Randrianasolo, L.D. Pyeatt, Q-learning: From computer network security to software security, in: 2014 13th International Conference on Machine Learning and Applications, 2014, pp. 257–262, http://dx.doi.org/10.1109/ICMLA.2014.47.

[121] M. Yousefi, N. Mtetwa, Y. Zhang, H. Tianfield, A reinforcement learning approach for attack graph analysis, in: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/ 12th IEEE International Conference on Big Data Science and Engineering, TrustCom/BigDataSE, 2018, pp. 212–217, http://dx.doi.org/10.1109/TrustCom/BigDataSE.2018.00041.

[122] Z.S. Stefanova, K.M. Ramachandran, Off-policy q-learning technique for intrusion response in network security, Int. J. Inf. Control Comput. Sci. 11.0 (4) (2018) http://dx.doi.org/10.5281/zenodo.1316524.

[123] G. Karatas, O. Demir, O. Koray Sahingoz, Deep learning in intrusion detection systems, in: 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT, 2018, pp. 113–116, http://dx.doi.org/10.1109/IBIGDELFT.2018.8625278.

[124] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Trans. Neural Netw. Learn. Syst. (2019) 1–20, http://dx.doi.org/10.1109/TNNLS.2018.2886017.

[125] J.J. Stubbs, G.C. Birch, B.L. Woo, C.G. Kouhestani, Physical security assessment with convolutional neural network transfer learning, in: 2017 International Carnahan Conference on Security Technology, ICCST, 2017, pp. 1–6, http://dx.doi.org/10.1109/CCST.2017.8167800.

[126] R. Vinayakumar, K.P. Soman, P. Poornachandran, Applying convolutional neural network for network intrusion detection, in: 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI, 2017, pp. 1222–1228, http://dx.doi.org/10.1109/ICACCI.2017.8126009.

[127] T. Kim, S.C. Suh, H. Kim, J. Kim, J. Kim, An encoding technique for CNN-based network anomaly detection, in: 2018 IEEE International Conference on Big Data, Big Data, 2018, pp. 2960–2965, http://dx.doi.org/10.1109/BigData.2018.8622568.

[128] L. Mohammadpour, T.C. Ling, C.S. Liew, C.Y. Chong, A convolutional neural network for network intrusion detection system, in: Proceedings of the Asia-Pacific Advanced Network, vol. 46, pp. 50–55.

[129] Y. Fu, F. Lou, F. Meng, Z. Tian, H. Zhang, F. Jiang, An intelligent network attack detection method based on RNN, in: 2018 IEEE Third International Conference on Data Science in Cyberspace, DSC, 2018, pp. 483–489, http://dx.doi.org/10.1109/DSC.2018.00078.

[130] S. Althubiti, W. Nick, J. Mason, X. Yuan, A. Esterline, Applying long short-term memory recurrent neural network for intrusion detection, in: SoutheastCon 2018, 2018, pp. 1–5, http://dx.doi.org/10.1109/SECON.2018.8478898.

[131] B. Abolhasanzadeh, Nonlinear dimensionality reduction for intrusion detection using auto-encoder bottleneck features, in: 2015 7th Conference on Information and Knowledge Technology, IKT, 2015, pp. 1–5, http://dx.doi.org/10.1109/IKT.2015.7288799.

[132] F. Farahnakian, J. Heikkonen, A deep auto-encoder based approach for intrusion detection system, in: 2018 20th International Conference on Advanced Communication Technology, ICACT, 2018, p. 1, http://dx.doi.org/10.23919/ICACT.2018.8323687.

[133] M. Yeo, Y. Koo, Y. Yoon, T. Hwang, J. Ryu, J. Song, C. Park, Flow-based malware detection using convolutional neural network, in: 2018 International Conference on Information Networking, ICOIN, 2018, pp. 910–913, http://dx.doi.org/10.1109/ICOIN.2018.8343255.

[134] Y. Wang, J. An, W. Huang, Using CNN-based representation learning method for malicious traffic identification, in: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science, ICIS, 2018, pp. 400–404, http://dx.doi.org/10.1109/ICIS.2018.8466404.

[135] T.T. Teoh, G. Chiew, Y. Jaddoo, H. Michael, A. Karunakaran, Y.J. Goh, Applying rnn and j48 deep learning in Android cyber security space for threat analysis, in: 2018 International Conference on Smart Computing and Electronic Enterprise, ICSCEE, 2018, pp. 1–5, http://dx.doi.org/10.1109/ICSCEE.2018.8538405.

[136] F. Meng, Y. Fu, F. Lou, A network threat analysis method combined with kernel PCA and LSTM-RNN, in: 2018 Tenth International Conference on Advanced Computational Intelligence, ICACI, 2018, pp. 508–513, http://dx.doi.org/10.1109/ICACI.2018.8377511.

[137] A.M. Vartouni, S.S. Kashi, M. Teshnehlab, An anomaly detection method to detect web attacks using stacked auto-encoder, in: 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems, CFIS, 2018, pp. 131–134, http://dx.doi.org/10.1109/CFIS.2018.8336654.

[138] D. He, D. Wang, Robust biometrics-based authentication scheme for multiserver environment, IEEE Syst. J. 9 (3) (2015) 816–823, http://dx.doi.org/10.1109/JSYST.2014.2301517.

[139] J.J. Hathaliya, S. Tanwar, S. Tyagi, N. Kumar, Securing electronics healthcare records in healthcare 4.0: A biometric-based approach, Comput. Electr. Eng. 76 (2019) 398–410, http://dx.doi.org/10.1016/j.compeleceng.2019.04.017.

[140] R. Gupta, S. Tanwar, S. Tyagi, N. Kumar, M.S. Obaidat, B. Sadoun, HaBiTs: Blockchain-based telesurgery framework for healthcare 4.0, in: 2019 International Conference on Computer, Information and Telecommunication Systems, CITS, 2019, pp. 1–5, http://dx.doi.org/10.1109/CITS.2019.8862127.

[141] K.A. Shakil, F.J. Zareen, M. Alam, S. Jabin, BAMHealthCloud: A biometric authentication and data management system for healthcare data in cloud, J. King Saud Univ. Comput. Inf. Sci. (2017) http://dx.doi.org/10.1016/j.jksuci.2017.07.001.

[142] M. Mohammedi, M. Omar, W. Aitabdelmalek, A. Mansouri, A. Bouabdallah, Secure and lightweight biometric-based remote patient authentication scheme for home healthcare systems, in: 2018 International Symposium on Programming and Systems, ISPS, 2018, pp. 1–6, http://dx.doi.org/10.1109/ISPS.2018.8379017.

[143] R. Gupta, S. Tanwar, S. Tyagi, N. Kumar, Tactile-internet-based telesurgery system for healthcare 4.0: An architecture, research challenges, and future directions, IEEE Netw. 33 (6) (2019) 22–29, http://dx.doi.org/10.1109/MNET.001.1900063.

[144] M. Wazid, A.K. Das, N. Kumar, M. Conti, A. Vasilakos, A novel authentication and key agreement scheme for implantable medical devices deployment, IEEE J. Biomed. Health Inf. PP (2017) 1, http://dx.doi.org/10.1109/JBHI.2017.2721545.

[145] P. Mehta, R. Gupta, S. Tanwar, Blockchain envisioned UAV networks: Challenges, solutions, and comparisons, Comput. Commun. 151 (2020) 518–538, http://dx.doi.org/10.1016/j.comcom.2020.01.023.

[146] C.G.L. Krishna, R.R. Murphy, A review on cybersecurity vulnerabilities for unmanned aerial vehicles, in: 2017 IEEE International Symposium on Safety, Security and Rescue Robotics, SSRR, 2017, pp. 194–199, http://dx.doi.org/10.1109/SSRR.2017.8088163.

[147] B. Yang, T. Zhang, A scalable meta-model for big data security analyses, in: 2016 IEEE 2nd International Conference on Big Data Security on Cloud, BigDataSecurity, IEEE International Conference on High Performance and Smart Computing, HPSC, and IEEE International Conference on Intelligent Data and Security, IDS, 2016, pp. 55–60, http://dx.doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.71.

[148] G. Papale, L. Sgaglione, SDD sentinel: A support tool for detecting and investigating electronic transaction frauds, in: 2017 IEEE International Conference on Internet of Things, IThings, and IEEE Green Computing and Communications, GreenCom, and IEEE Cyber, Physical and Social Computing, CPSCom, and IEEE Smart Data, SmartData, 2017, pp. 318–323, http://dx.doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.53.

[149] L. Vegh, Cyber-physical systems security through multi-factor authentication and data analytics, in: 2018 IEEE International Conference on Industrial Technology, ICIT, 2018, pp. 1369–1374, http://dx.doi.org/10.1109/ICIT.2018.8352379.

[150] A.A. Ahmed, L.M. Khay, Securing user credentials in web browser: Review and suggestion, in: 2017 IEEE Conference on Big Data and Analytics, ICBDA, 2017, pp. 67–71, http://dx.doi.org/10.1109/ICBDAA.2017.8284109.

[151] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, J.J. Rodrigues, Tactile internet for smart communities in 5G: An insight for NOMA-based solutions, IEEE Trans. Ind. Inf. (2019) 1, http://dx.doi.org/10.1109/TII.2019.2892763.