Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Anomaly Detection for Cyber Internet of Things Attacks: A Systematic Review

Laraib Sana, Muhammad Mohsin Nazir, Muddesar Iqbal, Lal Hussain & Amjad Ali

To cite this article: Laraib Sana, Muhammad Mohsin Nazir, Muddesar Iqbal, Lal Hussain & Amjad Ali (2022) Anomaly Detection for Cyber Internet of Things Attacks: A Systematic Review, Applied Artificial Intelligence, 36:1, 2137639, DOI: 10.1080/08839514.2022.2137639

To link to this article: https://doi.org/10.1080/08839514.2022.2137639

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2137639 (3263 pages)
https://doi.org/10.1080/08839514.2022.2137639

Taylor & Francis
Taylor & Francis Group

Anomaly Detection for Cyber Internet of Things Attacks: A Systematic Review

Laraib Sana[a], Muhammad Mohsin Nazir[a], Muddesar Iqbal[b], Lal Hussain[c,d], and Amjad Ali[e]

[a]Department of Computer Science, Lahore College for Women University, Jail Rd, near Wapda Flats, Jubilee Town, Lahore, Punjab; [b]Renewable Energy Laboratory, Communications and Networks Engineering Department, College of Engineering, Prince Sultan University, Riyadh, Saudi Arabia; [c]Department of Computer Science & IT, Neelum Campus, The University of Azad Jammu and Kashmir, Athmuqam, Azad Kashmir, Pakistan; [d]Department of Computer Science & IT, King Abdullah Campus, The University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir, Pakistan; [e]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan

ABSTRACT

From past few years, the Internet of things (IoT) is an emerging and encouraging technology that has gained prominence in the industries. Due to its increasing usages, a huge amount of data are exchanged within IoT architecture using the internet, which is why privacy and cyber-security are major issues. The heterogeneous nature of various technologies that are combined using IoT makes it problematic to provide security using prescriptive networking. The future of secure IoT depends on privacy issues. The research intends to improve security mechanisms based on intrusion and anomaly detection for IoT using deep learning. In this context, a systematic literature review (SLR) is conducted to identify 'How to perform data transformation analysis of IoT dataset to detect anomaly detection for cyber IoT attacks? The SLR result found 24 datasets used for IoT analysis, 35 performance metrics to evaluate IoT problems, 6–42 features identified for detection, 42 preprocessing techniques have been used for transforming data, and 26 different methods and models were used to process the given problem. The SLR highlights further enhancement for the issue and identification of cyber-security in IoT. Anomaly detection can be done based on reinforcement deep learning after a thorough analysis of SLR.

ARTICLE HISTORY
Received 4 June 2022
Accepted 13 October 2022

Introduction

Internet of Things (IoT) can be referred to as a mixture of physical devices such as same digital devices, i.e. sensors, actuators, routers, smartphone and all those devices which can be embedded with software and computer-based system (Ray 2018) (Das, Zeadally, and He 2018). The technology of IoT allows the collection of data from different physical and digital devices from distributed network infrastructure and makes them communicate over internet for

CONTACT Muddesar Iqbal miqbal@psu.edu.sa Renewable Energy Laboratory, Communications and Networks Engineering Department, College of Engineering, Prince Sultan University, Riyadh 11586, Saudi Arabia

processing purposes (Diro and Chilamkurti 2018). This feature has made the IoT technology suitable for the real-world application in which there is a need to merge physical devices with computer-based system such as smart cities, smart health, smart homes, smart parking, industrial internet, energy management, smart agriculture, etc. (Fisher et al. 2018). These devices can communicate this information to internet through trusted gateway nodes, and this information can also be accessed by users, smart e-health, smart grids, smart cities, smart homes, etc. (Das, Zeadally, & He, 2018). But the complexity of unknown cyber-attacks has raised questions on the adoption of these smart devices (Das, Zeadally, and He 2018).

Usually, an IoT system has three layers that are known as physical perception layer, network layer and an application layer (Čolaković and Hadžialić 2018). Physical perception layer deals with the sensor, actuators and environment, and it perceives information from the environment and human, collects all the information and converts all the data in a format that is acceptable in cyber world (Skarmeta et al. 2014). Network layer includes all the network components such as wireless sensor devices, nano-devices, cellular network and ad hoc network in heterogeneous network for data processing (coding, transmission, fusion, analysis) to produce essential information that can be accessed by application layer and application layer to further offer services to IoT end users (Sicari et al. 2015). All sensors in IoT are defenseless against attacks and threats. The massive network traffic is challenging to implement the intrusion detection system due to demand of better detection efficiency with the goal that attack can be identified continuously. But due to limitation and constraints, for example, late alerts, high false alert rate and low recognition effectiveness, such detection system fails, which is why it is difficult to maintain security in IPV6/RPL connected network (Grammatikis, Sarigiannidis, and Moscholios 2019).

Intrusions are the main reason that can cause hurdles for the adoption of IoT infrastructure. An intrusion detection system is composed of different intrusion principles and mechanisms over internet (Raza, Wallgren, and Voigt 2013). Intrusion detection techniques consists of software and hardware that can be implemented on any host node. It can monitor any suspicious activities and traffic coming from nearby nodes. By observing these traffics, malicious nodes and cyber-attacks can be identified (Santos et al. 2018). The goal of IDS is to inform the managers about malicious threats and to protect the network from unpredictable attack (Elrawy, Awad, and Hamed 2018). Intrusion detection system includes signature-based strategy, anomaly and specification-based approaches (Elazhary 2019). In signature-based detection, incoming traffic is matched with the already-known attacks stored in database, while in anomaly detection, attacks can be detected by deviation in their normal behavior. To develop the normal behavior profile, researchers used statistical techniques and machine learning. In specification-based detection, a specification is maintained that can be called rules of normal behaviors, and detects intrusions if any connected device deviates from their normal behavior (Diro and Chilamkurti 2018). In the literature, many machine learning (ML) and deep learning (DL) techniques have been adopted for developing NIDS (Sharma et al. 2017). Statistical, machine-learning- and deep-learning-based mechanisms are utilized in the anomaly detection. There are four basic types of machine learning algorithm: supervised, unsupervised, semi supervised and reinforcement learnings (Xiao et al. 2018). The problem with supervised machine learning algorithm is that it requires label data. Some well-known intrusion detection mechanisms based on supervised learning such as support vector machine, convolutional neural network and auto encoder have been developed (Jagannath et al. 2019). Research shows that unsupervised machine learning can provide better accuracy than supervised learning. The most used unsupervised machine learning algorithm for intrusion detection is clustering (Hosseinpour et al. 2016). A combination of supervised learning and unsupervised learning known as semi-supervised learning has also been used for the purpose of intrusions detection. Fuzzy C-Means is one example of such semi-supervised learning algorithm (Zeng et al. 2013). However, the prediction of any learning algorithm is totally dependent on training and testing dataset, and preprocessing of dataset is a crucial part of any learning algorithm.

Prediction of any learning algorithm is totally dependent on training and testing dataset, and preprocessing of dataset is a crucial part of any learning algorithm. This study mainly deals with data transformation analysis for anomaly detection of IoT datasets. Systematics literature review (SLR) will be conducted to perform the data transformation analysis. SLR is a systematic way to identify and evaluate the available research against research question. With the help of SLR, we will figure out the preprocessing techniques, methods/model, features and available datasets and evaluation parameters for anomaly detection of IoT datasets.

(1) Techniques
(2) Methods
(3) Features
(4) Selection of dataset
(5) Performance metrics

Research Question: How to perform data transformation/preprocessing analysis of IoT datasets to perform anomaly detection for cyber IoT attacks?

Methodology

This methodology is carried out by means of SLR presented by Gupta et al. (2019), Kitchenham (2004) and Zahra et al. (2017). The methodology consists of five steps such as review protocol development, inclusion and exclusion criteria, search process, quality assessment, data extraction and synthesis. In review protocol development, we specify the methods to carry out the SLR. Review protocol helps us in selection and further analysis of research/journal papers. It will include background, research question, search strategy and study selection criteria. Inclusion and exclusion criteria describe which things will be included or excluded in the search. Search process will describe the overall search process and search flow. Quality assessment will describe the criteria used to assess the quality of SLR. In data extraction, a strategy will be defined from which data are extracted for primary study. Data synthesis will help to identify whether meta-analysis is required or not and if it is required, what strategy will be used.

Review Protocol Development

Inclusion and Exclusion Criteria

The following inclusion and exclusion criteria are used to carry out the systematic literature review as reflected in Table 1.

(1) Only those papers have been selected that contain the keywords shown in Table 1. All others papers have been excluded.
(2) Only those researches have been considered that have been published during 2010–2020 in IEEE, Elsevier, Springer, ACM and Taylor & Francis. All papers from other journal papers have been excluded.
(3) Journal papers and chapters are only considered. Conferences, books and magazines have been excluded.
(4) Papers written in English are included; all papers written in other language have been excluded.
(5) Publications must have data preprocessing techniques or methods, features, for IoT datasets and evaluation parameters for anomaly detection.

Table 1. Search terms and their results.

Sr.	Search Terms	Operators	IEEE	Springer	ACM	Science Direct	Taylor & Francis
1	Data Preprocessing	AND	1299	10,366	476	14,584	**1214**
2	Data Preprocessing IoT	AND	30	462	41	1178	35
3	Data Preprocessing on IoT Datasets	AND	6	319	24	751	21
4	Data Preprocessing on IoT Datasets Anomaly	AND	1	75	7	229	6
5	Data Preprocessing on IoT Datasets for Anomaly detection	AND	1	73	7	197	6

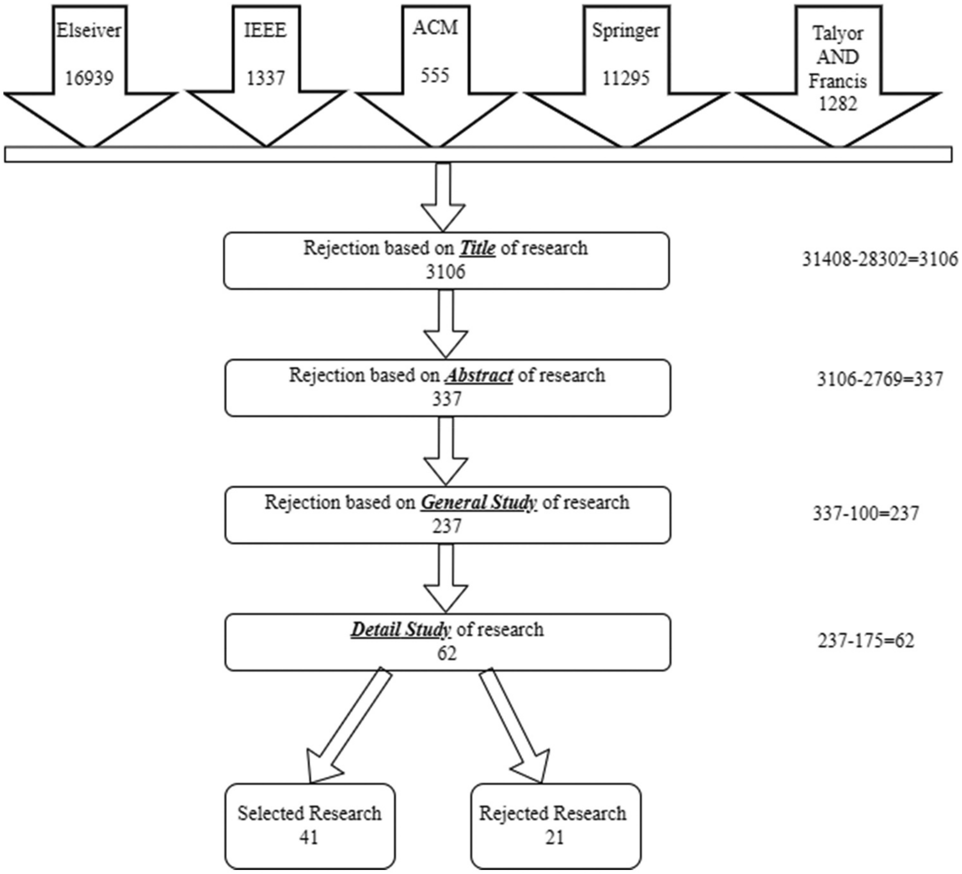

Figure 1. Search process.

Search Process

The search process is carried out through keywords shown in Table 1. The whole procedure of searching process is mentioned in Figure 1 and Figure 4.

Quality Assessment

Quality assessment is necessary for the quality authentication of the proposed study. Quality assessment is an aspect of the rigorous formative evaluation process that involves reliability and efficiency. In order to bring accuracy in our results, we have maintained the required quality standard.

For quality assessment of the proposed study, the latest articles from reputed scientific databases have been selected.

Our selected researches are based on realistic approach.

Table 2. Data extraction.

Sr.	Description	Details
1	Criteria	Title, year, type (journal), abstract, publisher
Extraction of Data		
Overview	Whether aims and goals of our selected study are according to our criteria?	
Methods	Which qualitative and quantitative methods are used in selected study?	
Results	What are the results of selected study?	
Datasets	Which datasets are used?	
Validation	Validate the techniques from selected researches.	
Synthesis of Data		
Techniques	Which techniques are used for data preprocessing of IoT datasets?	

Summary.

Data Extraction and Synthesis

We have extracted the data from the identified 41 research papers. The selection criteria for data extraction and synthesis are mentioned in Table 2.

Publication Based on Method/Model

D. Singh et al. have raised the importance of human behavior analysis in the field of smart environment (Kaur et al. 2019). A model using binary cuckoo search-based metaheuristic framework is proposed that has used multiple base learners to recognize human activities. However, the research can be improved by optimizing the classifier parameters based on metaheuristic techniques (Kaur et al. 2019). A. Tolba et al. have presented the heart disease predictive model based on internet of medical things and high-order Boltzmann model (Al-Makhadmeh and Tolba 2019). Results showed that HOBDBNN provided better accuracy (99.03%) compared to other classifiers and predicted the abnormal heart patterns from a large number of datasets in a minimum time (8.5 s) by reducing analytical complexity. S. Huang et al. have proposed a model based on deep learning for enhancing the performance of content delivery network (CDN) (Zhang et al. 2018). Recurrent neural network has used long short-term memory (LSTM) to model the temporal data in different sequence time. However, in future, more cache groups must be considered to validate the proposed model. The authors have proposed a model based on deep learning using long-term memory for flight landing prediction system (Tong et al. 2018). The results showed that the proposed model performed better in the prediction of flight decision and air traffic controllers for air traffic management. Stochastic compression technique has been designed to handle adversarial attack for DNN-based voice-enabled systems. Results showed that the use of standard cross entropy and adversarial losses in training helped in the effectiveness of the proposed model (Bhattacharya et al. 2020). An energy-efficient preprocessing and anomaly detection method has been discussed that is based on long short-term memory network. A case analysis of manufacturing system was used to implement the proposed method. Results showed that the method performed better anomaly detection (Zhang and Ji 2020). A semi-supervised model based on CNN and generic adversarial network was proposed. The study performed three comparisons: the first comparison was done with modern GAN-Based Models and other comparisons were with Ablated Fully Supervised Benchmarks and state-of-the-art models. Results showed that their model performed better (Wang et al. 2019a). An anomaly detection method based on recurrent LSTM-based autoencoder and convolution neural network has been proposed. The study has also addressed the issues of anomaly detection for univariable time series data. In future, this method can be used for multivariable time series IoT data (Demertzis et al. 2020). In this paper, an anomaly detection method based on multi-classification deep learning and LSTM has been presented for handling time series datasets. Results showed that the proposed method worked better with concept drift method and multi classification method (Xu et al. 2020). In this paper, resource constraint of IoT devices had been handled by introducing distributed and semi-distributed cases for intrusion detection. Parallel machine-learning-based method had been used for feature selection and extraction. Results of comparative analysis showed that the detection accuracy of the proposed method with semi-distributed case was better than other case (Rahman et al. 2020).

Publication Based on Framework

A health framework has been designed based on IoT and ubiquitous computing to predict probabilistic vulnerability at person's workplace (Bhatia and Sood 2017). S. K. Sood et al. have used framework based on artificial neural network to predict intelligently. The prediction model consisted of three layers: monitor, learn and predict. The authors have proposed the framework based on two machine learning algorithms: graph-based clustering algorithm and Bayesian probabilistic graphical model (Sun et al. 2019). Multivariate Hawkes process had been used to control the attackers' information. The proposed method is better than other methods as it does not require any prior knowledge. F. Arif et al. have proposed an architecture based on big data analytics and Hadoop server for effective planning and decision-making in smart cities (Aljawarneh and Vangipuram 2018). The implementation of the proposed architecture was done by using Hadoop server with map reduce mechanism and java programming. Pre-processing with min-max techniques, rules engine and filtration made the datasets more suitable for Hadoop server by reducing pre-processing computation at server side. A framework for energy management in controllable IoT environment has been proposed. The idea behind their research was to predict the future energy demand fluctuations using a prediction algorithm and to make the communication easier between consumer and producer. In future, model can be modified for IoT devices for mutual energy sharing to save energy consumption (Han et al. 2020). This paper has proposed a framework named ADE based on machine learning for detection of denial-of-service attack DoS. Multi-scheme AD1E, AD2E was also compared with another classifier. Results showed that detection rate of AD2E with MLP was accurate than another classifier (Baig et al. 2020).

Publication Based on Security System

For efficient intrusion detection system, a novel technique GARUDA based on distance measure had been proposed for feature selection and feature representation (Aljawarneh and Vangipuram 2018). The major contribution of this study is to provide optimal observation matrix by applying incremental future clustering. In this study, the authors have proposed a security mechanism based on deep learning for social fog IoT (Diro and Chilamkurti 2018). The research indicates that self-learning and compression techniques make the deep learning suitable for distributed attack detection mechanism. The results showed that deep learning took more time in learning as compared to traditional machine learning; however, the detection rate of both techniques was same. The authors have presented the intrusion detection scheme for IoT architecture based on naive Bayesian classification algorithm (Mehmood et al. 2018). They have used multi-agents to sense the abnormalities and irregularity in the traffic and IoT nodes. This study has improved the security in network layer by handling distributed denial service or attack. In future, this study can be extended by designing lightweight machine learning algorithm for IDS. A target search system called Target Finder, a privacy-preserving protocol and a work selection problem have been designed and implemented using an Android app. Light CNN has been used for feature extraction. Real-life implementation provided better results in privately locating targets (Khazbak et al. 2020). In this paper, the authors have provided the three-layer security mechanism based on MLTP and Legendre approximation for implantable medical devices. The proposed security mechanism deals with data layer, network layer and application layer. For authentication, ECG signals have been used. In future, this technique can be used for other types of datasets (Rathore et al. 2020).

Publication Based on Preprocessing Techniques

The main concern is missing data due to error in recording through instruments with respect to medical IoT datasets (Fisher et al. 2018). This study has mainly focused on techniques based on a Markov model for handling such missing data. The results showed that the proposed model would work better in the prediction of missing value in the datasets. S. P. Sonavanebhas et al. have proposed an IoT-based smart farming environment using an improved genetic algorithm that is based on multi-level parameter feature selection algorithm (Kale and Sonavane 2019). ELM classifiers had been used that improved the classification by reducing the number of features for classification. The proposed model is only evaluated for binary classification, and it should also be done for multi-class-level classification in future. Y. Tsao et al. have analyzed and proposed a procedure based on deep learning and multi-style learning using denoising autoencoder (DAE) for preprocessing (Lin et al. 2017). The study showed that combining synthesized and original training datasets can improve the training results of DNNs classifier and enhance the capability of deep learning for data preprocessing and collection. In another study, the authors have presented the idea of error and event in the outliers (Nesa, Ghosh, and Banerjee 2018). The authors had proposed a learning-based mechanism to handle error and event in outlier detection for IoT environment. Exhaustive cross-validation method has been used for validation of the proposed model. Results showed that the proposed model works better for outlier detection compared to other state-of-the-art works. The authors have proposed a fog-based distribution attack detection using ELM classifier and semi-supervised Fuzzy C-Means (ESFCM) method to handle the issue of scalability, distribution, resource limitations and low latency (Rathore and Park 2018). ELM classifier algorithm is designed based on feed forward neural network with single hidden layer. The performance of framework had also been evaluated by comparing with centralized cloud-based framework that showed 86.53% accuracy in attack detection. The authors have not dealt with the issue of random assignment of input bias and weights for ELM that can result in lower performance. This issue can be solved by using deep learning technique for better attacks detection. The authors have discussed the design of intrusion detection system that involved three major issues: dimensionality reduction, distance measure and selection of learning classifier (Aljawarneh and Vangipuram 2018). For efficient intrusion detection system, a novel technique GARUDA based on distance measure had been proposed for feature selection and feature representation. The GARUDA-based approach had improved the detection rate accuracy up to 99.78%. The experiments showed that the proposed approach based on GARUDA measure had better results for classification and detection accuracies compared to other classifiers for low-frequency U2 R (User to Root attack) and R2 L (Root to Local attacks) attack classes. kNN and J48 classifier had performed best for U2 R and R2 L attack classes; however, GAURDA approach had failed to improve the accuracies of the classification and detection rates of SVM classifier for U2 R and R2 L attack classes. A baseline suppression technique has been introduced to handle the interference problem in e-nose of IoT. Median filtering and standardization have been used for preprocessing. Principal component analysis (PCA) and independent component analysis (ICA) have been used for dimensionality reduction (Wang et al. 2019c). This paper has discussed the issue of missing values with larger gaps for sensor data. They have discussed the importance of preprocessing for IoT infrastructure. An imputation technique has been proposed to handle the missing values for univariate time series data in industrial based IoT system.. The proposed method performed better than state-of-the-art methods in handling large gaps missing sensor data. However, only sensor data were evaluated (Liu et al. 2020). This paper is mainly concerned with preprocessing issues such as privacy and delay in data collection for the internet of vehicles. The purpose of this scheme was to minimize the error in data by training a model based on edge computing. Semi-supervised based deep learning and convolution neural networks have been used to find a correlation between required data and images. Study showed that delay in the proposed scheme is lower than the earlier scheme based on supervised learning and reduced the amount of uploading data on the cloud (Wang et al. 2019b).

Other Publications

Sarker and Salah (2019) have proposed a concept of modeling context-aware smartphone apps using machine learning prediction techniques such as random forest learning. For preprocessing, label encoding technique has been used to convert the categorical contextual data into discrete data. A tool named IoT inspector has been introduced to gather real-life labeled network dataset from home devices. During this crowdsourcing, several security and privacy issues have been identified. This tool provides better-quality real-life datasets of smart homes comparatively from lab-created datasets (Huang et al. 2020).

Results

The publications based on leading journals are reflected in Figure 2.

The studies based on anomaly detection in IoT are covered in a wide range of well-known journals. Twenty-two journals have been identified in SLR. Future Generation Computer Systems with 17.07% is at the top first. The Journal of Super Computing, Internet of Things and IEEE Transactions on Industrial Informatics are on the second and third positions, respectively. Their percentages of publication are, respectively, 9.76% and 7.32%. Other journals have 4.88% and 2.44% with 2 and 1 publications, respectively. From the SLR, we can conclude that Future Generation Computer Systems is most feasible for the publication related to anomaly detection in IoT.

Publication based on scientific database is reflected in Figure 3.

SLR was conducted from five scientific databases which are ACM, Elsevier, IEEE, Springer and Taylor & Francis. Most papers selected for SLR studies are from Elsevier with 21 papers. IEEE and Springer have the same number of papers which is eight papers. The least number of selected papers is from Taylor & Francis. During the search, it was identified that most relevant papers

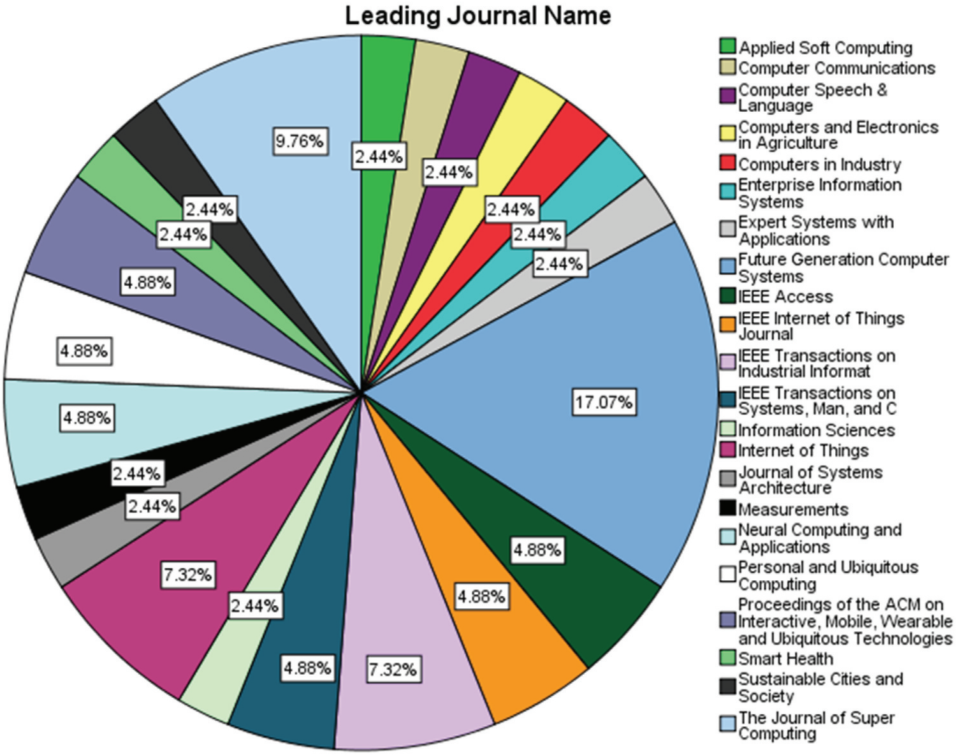

Figure 2. Publication based on leading journals.

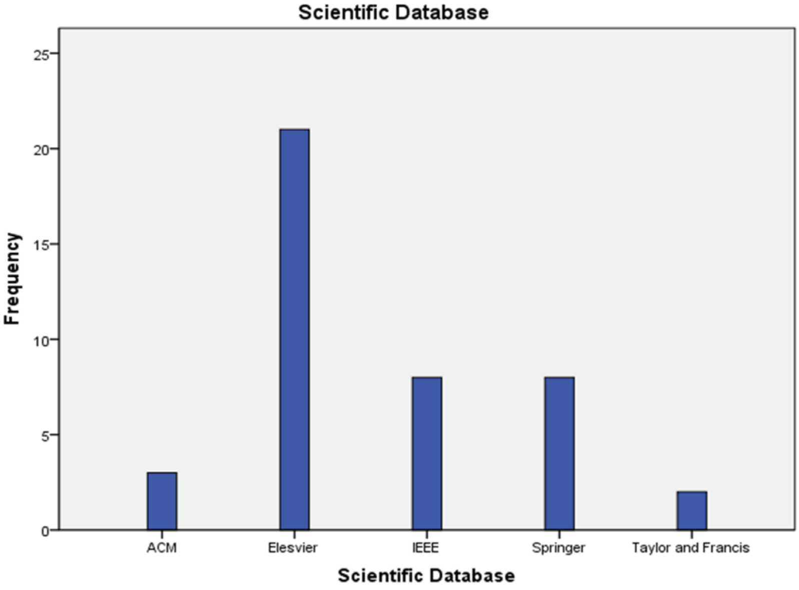

Figure 3. Publication based on scientific database.

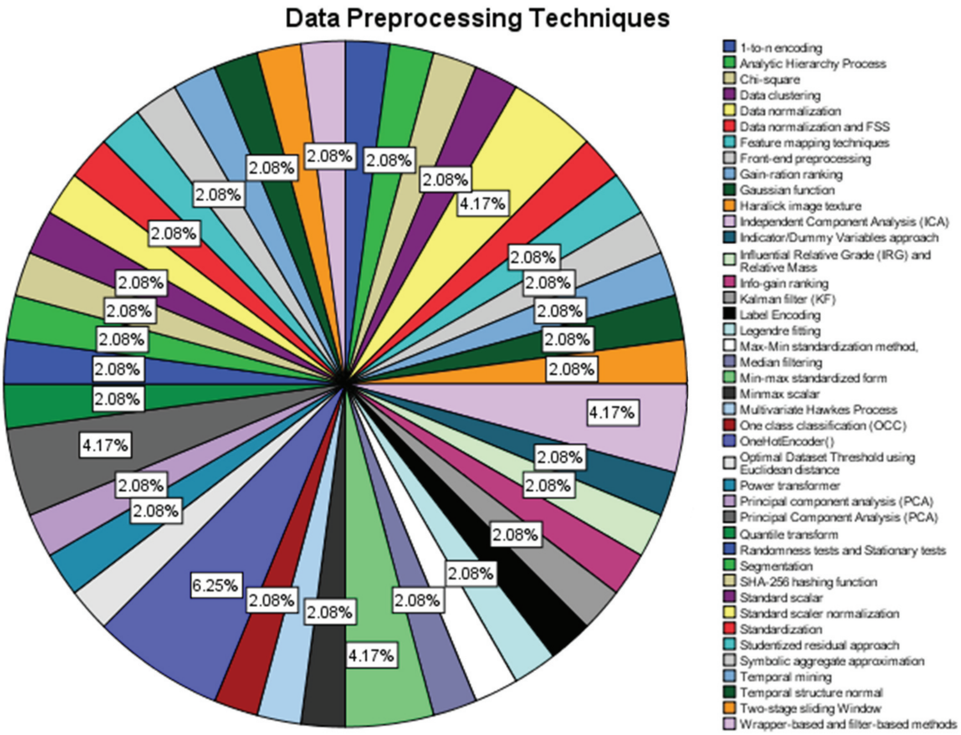

Figure 4. Data preprocessing technique.

related to anomaly detection are found in Science Direct/Elsevier. It is concluded that IoT researchers prefer Elsevier for their publications.

Technique

The data preprocessing methods are detailed in the Figure 4.

In SLR, 42 techniques have been identified for preprocessing of IoT datasets. The most used preprocessing technique is OneHotEncoder() with a percentage of 6.25%. Other most used preprocessing techniques are data normalization, independent component analysis, min-max standardized form and principal component analysis with a percentage of 4.17%. Remaining 37 preprocessing techniques have the same percentage, 2.08%.

Datasets of IoT

The extracted datasets of IoT are depicted in Figure 5. From SLR studies, 24 datasets have been identified. NSL KDD CUP is the mostly used dataset with a percentage of 14.29% among all datasets. Another commonly used dataset is medical datasets with a percentage of 11.43%. It was also observed during the

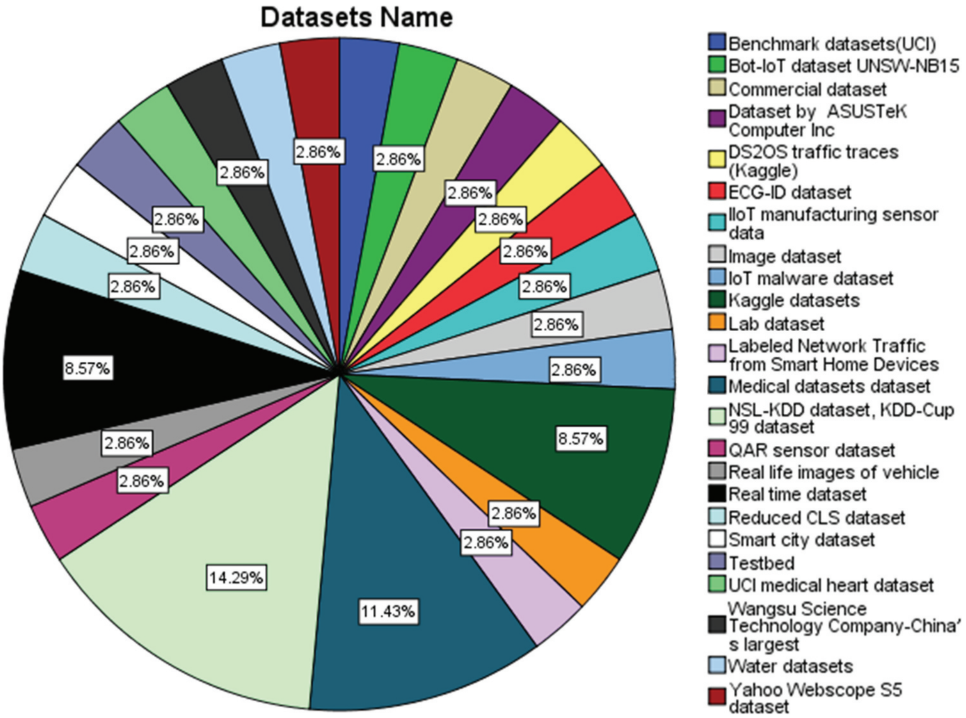

Figure 5. Extracted datasets of IoT.

study that mostly researchers have used real-life datasets from their experiments. Kaggle dataset is also used by many IoT researchers. Percentage of real-life dataset and Kaggle datasets is 8.57%. Remaining datasets have the same percentage of 2.86% among the results.

Model/Method

Figure 6 shows the extracted models/methods. A total of 26 methods/framework/model have been identified in SLR detailed studies. The most common methods used for anomaly detection for IoT are artificial neural network and deep learning based with percentages of 12.12% and 9.09%, respectively. Multi-layer perceptron model is also used in two papers, with a percentage of 6.06%. Results showed that deep-learning-based model is more feasible for IoT environment.

Performance Metrics

Figure 7 indicates the extracted performance metrics of IoT. A total of 35 performance metrics have been figured out from the SLR studies. The most used performance metrics is accuracy and F score with a percentage of 14.46%.

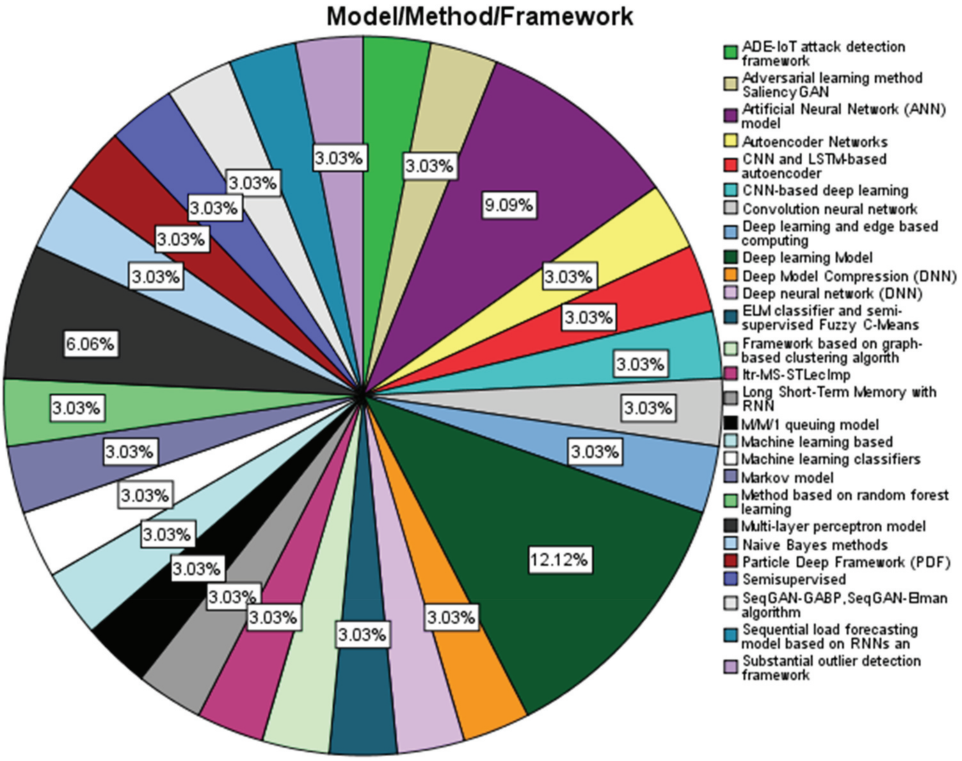

Figure 6. Extracted model/method.

Precision and recall are also commonly used, having percentages of12.05% and 9.64%, respectively, among the results. Detection rate is also popular for evaluation metric with a percentage of 3.61%. False alarm rate, ROC curve, sensitivity, specificity and normalization have the same percentage, 2.41%.

Number of Features

There are six papers in research study, which contains information about features, and 6–42 features are identified.

Table 3 indicates the number of selected features against research identification.

Discussion and Limitations

Table 4 reflects the relationship of identification research with respect to SLR.

This study presents a systematic literature on anomaly detection issues for cyber IoT. We have selected and reviewed 41 papers till 2020. The results showed that the first 8 years did not have any specific publication related to IoT. Mostly, publication was done between 2017 and 2020. Twenty-two journals have been

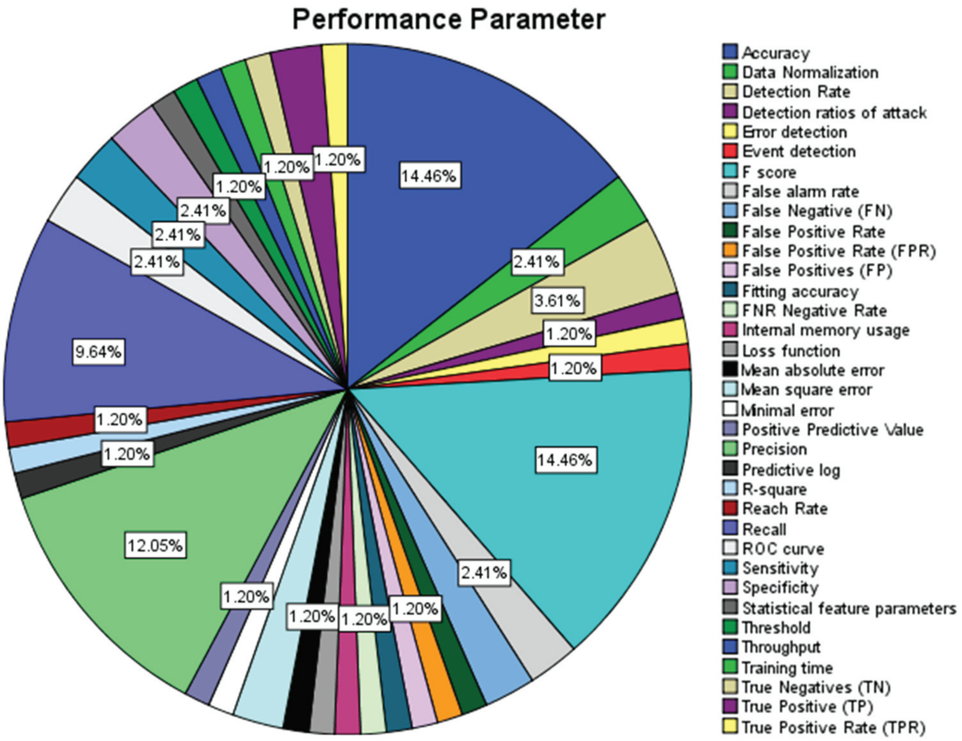

Figure 7. Extracted performance metrics of IoT.

Table 3. Number of selected features against research identification.

Sr.	No. of Selected Features	Research Identification
1.	6 features	Kaur et al. (2019)
2.	19,41 features	Aljawarneh and Vangipuram (2018)
3.	41 features	Bhatia and Sood (2017)
4.	13 features	Fisher et al. (2018)
5.	9 features	Hasan et al. (2019)
6.	42 features	Rathore and Park (2018)
7.	20 features	Sun et al. (2019)

identified in SLR that are Future Generation Computer Systems, Journal of Supercomputing, Internet of Things and IEEE Transactions on Industrial Informatics. Our study has also identified that (n = 21) papers of SLR belong to Science Direct/Elsevier. Our study showed that OneHotEncoder (6.25%) is the mostly used preprocessing techniques among 42 preprocessing techniques, and accuracy and f-score is the most used performance metric among 34 performance metrics. Twenty-four datasets have been identified through SLR. It revealed that the use of real-life IoT datasets is popular among researchers. However, Bot-IoT dataset UNSW-NB15 has potential and can be used for designing security mechanism for IoT environment. Table 4 shows the model implementation with respect to anomaly detection for cyber IoT. Twenty-six models are checked against preprocessing technique, performance metric and

Table 4. Relationship of identification research with respect to SLR.

Sr.	Model/Method	Research Identification	Preprocessing Techniques	Performance Metric	Dataset
1.	Machine learning based	Kaur et al. (2019)	✓	✓	✓
2.	M/M/1 queuing model, Map Reduce Mechanism	Babar and Arif (2017)	✓		✓
3.	Artificial Neural Network (ANN) model	Al-Makhadmeh and Tolba (2019); Bhatia and Sood (2017), Kale and Sonavane (2019)	✓ ✓ ✓	✓ ✓ ✓	✓
4.	Deep learning Model	Diro and Chilamkurti (2018), Zhang et al. (2018); Tong et al. (2018); Lin et al. (2017)	✓ ✓ ✓ ✓	✓ ✓	✓ ✓ ✓ ✓
5.	Markov model	Fisher et al. (2018)		✓	✓
6.	Naive Bayes methods	Mehmood et al. (2018)	✓	✓	✓
7.	Substantial outlier detection framework	Nesa, Ghosh, and Banerjee (2018)	✓	✓	✓
8.	ELM classifier and semi-supervised Fuzzy C-Means	Rathore and Park (2018)	✓	✓	✓
9.	Method based on random forest learning	Sarker and Salah (2019)	✓	✓	✓
10.	Framework based on graph-based clustering algorithm and bayesian probabilistic graphical model	Sun et al. (2019)	✓		✓
11.	Deep learning and edge-based computing	Wang et al. (2019b)		✓	✓
12.	*Sequential load forecasting model based on RNNs and LSTM*	Han et al. (2020)	✓	✓	✓
13.	CNN and LSTM-based autoencoder	Yin et al. (2020)	✓	✓	✓
14.	Semisupervised adversarial learning method SaliencyGAN	Wang et al. (2019a)		✓	✓
15.	Deep Model Compression (DNN)	Bhattacharya et al. (2020)			
16.	SeqGAN-GABP, SeqGAN-Elman algorithm	Geng and Du (2020)	✓	✓	
17.	Autoencoder Networks	Demertzis et al. (2020)	✓	✓	
18.	Multi-layer perceptron model	Hosseinzadeh et al. (2020)	✓	✓	✓
19.	Long Short-Term Memory with RNN	Xu et al. (2020)			
20.	ADE-IoT attack detection framework	Baig et al. (2020)	✓	✓	✓
21.	Machine learning classifiers	Karanja, Masupe, and Jeffrey (2020)	✓	✓	✓
22.	Deep neural network (DNN)	RM et al. (2020)	✓	✓	
23.	CNN-based deep learning	Jung et al. (2020)	✓	✓	✓
24.	Particle deep framework (PDF)	Koroniotis, Moustafa, and Sitnikova (2020)	✓	✓	✓
25.	Itr-MS-STLecImp	Liu et al. (2020)			✓
26.	Convolution neural network	Shi et al. (2020)		✓	✓

datasets. The table shows which model had used which technique, metric and dataset. Fourteen papers are those that contain information about method, preprocessing technique, performance metric and datasets. There is not one paper in which only method has been discussed. Other papers have information about either preprocessing, datasets or performance metric. Our results also showed that deep-learning-based model is most popular for anomaly detection for cyber IoT.

Limitation. Although it has been tried best to do concise, consistent and correct systematic literature review, there are chances of missing some relevant studies.

Answer to RQ.

Research Question: How to perform data transformation/preprocessing analysis of IoT dataset to perform anomaly detection for cyber IoT attacks?

A total of 41 researchers have been identified by doing SLR for the period 2010–2020 based on exclusion and inclusion criteria. After detailed searching and study of each publication, we have identified 26 models, 42 preprocessing techniques, 35 performance metrics, 24 datasets and 6–42 features. Graphs are also made in SPS based on these findings.

Conclusion and Future Work. As this is the world of artificial intelligence and ubiquitous computing, the usage of IoT sensor is prevailing in every aspect of life. Now, we have smart homes, smart agriculture, smart medical, smart city, etc. As we are incorporating IoT sensors in our daily life, the security and privacy of every individual must be ensured. In this study, an SLR has been conducted for the period 2010–2020. The purpose of this SLR is to conduct data transformation analysis for anomaly detection of cyber IoT. Forty-one papers have been selected for detailed searching and study. This study identified 26 models, 42 preprocessing techniques, 35 performance metrics, 24 datasets and 6–42 features. From the result, it is concluded that deep-learning-based model is encouraging to use for anomaly detection method in IoT environment. It is also concluded that accuracy and OneHotEncoder are widely used performance metrics and preprocessing techniques, respectively, for IoT environment. In future, the identified model can be deeply analyzed for cyber IoT.

Acknolwedgements

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

Disclosure Statement

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

References

Aljawarneh, S. A., and R. Vangipuram. 2018. GARUDA: Gaussian dissimilarity measure for feature representation and anomaly detection in Internet of things. *The Journal of Supercomputing* 76 (6):1–38. doi:10.1007/s11227-018-2397-3.

Al-Makhadmeh, Z., and A. Tolba. 2019. Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: A classification approach. *Measurement* 147:106815. doi:10.1016/j.measurement.2019.07.043.

Babar, M., and F. Arif. 2017. Smart urban planning using big data analytics to contend with the interoperability in internet of things. *Future Generation Computer Systems* 77:65–76. doi:10.1016/j.future.2017.07.029.

Baig, Z. A., S. Sanguanpong, S. N. Firdous, T. G. Nguyen, and C. So-In. 2020. Averaged dependence estimators for DoS attack detection in IoT networks. *Future Generation Computer Systems* 102:198–209. doi:10.1016/j.future.2019.08.007.

Bhatia, M., and S. K. Sood. 2017. A comprehensive health assessment framework to facilitate IoT-assisted smart workouts: A predictive healthcare perspective. *Computers in Industry* 92:50–66. doi:10.1016/j.compind.2017.06.009.

Bhattacharya, S., D. Manousakas, A. G. C. P. Ramos, S. I. Venieris, N. D. Lane, and C. Mascolo. 2020. Countering acoustic adversarial attacks in microphone-equipped smart home devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2):1–24. doi:10.1145/3397332.

Čolaković, A., and M. Hadžialić. 2018. Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues. *Computer Networks* 144:17–39. doi:10.1016/j.comnet.2018.07.017.

Das, A. K., S. Zeadally, and D. He. 2018. Taxonomy and analysis of security protocols for Internet of Things. *Future Generation Computer Systems* 89:110–25. doi:10.1016/j.future.2018.06.027.

Demertzis, K., L. Iliadis, N. Tziritas, and P. Kikiras. 2020. Anomaly detection via blockchained deep learning smart contracts in industry 4.0. *Neural Computing & Applications* 32 (23):17361–78. doi:10.1007/s00521-020-05189-8.

Diro, A. A., and N. Chilamkurti. 2018. Distributed attack detection scheme using deep learning approach for internet of things. *Future Generation Computer Systems* 82:761–68. doi:10.1016/j.future.2017.08.043.

Elazhary, H. 2019. Internet of Things (IoT), mobile cloud, cloudlet, mobile IoT, IoT cloud, fog, mobile edge, and edge emerging computing paradigms: Disambiguation and research directions. *Journal of Network and Computer Applications* 128:105–40. doi:10.1016/j.jnca.2018.10.021.

Elrawy, M. F., A. I. Awad, and H. F. A. Hamed. 2018. Intrusion detection systems for IoT-based smart environments: A survey. *Journal of Cloud Computing* 7 (1):21. doi:10.1186/s13677-018-0123-6.

Fisher, P. S., J. James, J. Baek, and C. Kim. 2018. Mining intelligent solution to compensate missing data context of medical IoT devices. *Personal and Ubiquitous Computing* 22 (1):219–24. doi:10.1007/s00779-017-1106-1.

Geng, T., and Y. Du. 2020. The business model of intelligent manufacturing with internet of things and machine learning. *Enterprise Information Systems* 1–19.

Grammatikis, P. I. R., P. G. Sarigiannidis, and I. D. Moscholios. 2019. Securing the internet of things: Challenges, threats and solutions. *Internet of Things* 5:41–70. doi:10.1016/j.iot.2018.11.003.

Gupta, S. K., A. Gunasekaran, J. Antony, S. Gupta, S. Bag, and D. Roubaud. 2019. Systematic literature review of project failures: Current trends and scope for future research. *Computers & Industrial Engineering* 127:274–85. doi:10.1016/j.cie.2018.12.002.

Han, T., et al. 2020. An efficient deep learning framework for intelligent energy management in IoT networks. *IEEE Internet of Things Journal*. 8 5: 3170–3179. 10.1109/JIOT.2020.3013306

Hasan, M., M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem. 2019. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet of Things*. 7:100059. doi:10.1016/j.iot.2019.100059.

Hosseinpour, F., P. V. Amoli, J. Plosila, T. Hämäläinen, and H. Tenhunen. 2016. An intrusion detection system for fog computing and IoT based logistic systems using a smart data approach. *International Journal of Digital Content Technology and Its Applications*. 10:34–46 .

Hosseinzadeh M, Ahmed OH, Ghafour MY, Safara F, Ali S, Vo B, Chiang HS. 2020. A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. *The Journal of Supercomputing*. 2021 Apr;77(4):3616–37.

Huang, D. Y., N. Apthorpe, F. Li, G. Acar, and N. Feamster. 2020. Iot inspector: Crowdsourcing labeled network traffic from smart home devices at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2):1–21.

Jagannath, J., N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia. 2019.Machine learning for wireless communications in the internet of things: A comprehensive survey. *Ad Hoc Networks* 93: 101913.doi: 10.1016/j.adhoc.2019.101913.

Jung, W., H. Zhao, M. Sun, and G. Zhou. 2020. IoT botnet detection via power consumption modeling. *Smart Health* 15:100103. doi:10.1016/j.smhl.2019.100103.

Kale, A. P., and S. P. Sonavane. 2019. IoT based smart farming: Feature subset selection for optimized high-dimensional data using improved GA based approach for ELM. *Computers and Electronics in Agriculture* 161:225–32. doi:10.1016/j.compag.2018.04.027.

Karanja, E. M., S. Masupe, and M. G. Jeffrey. 2020. Analysis of internet of things malware using image texture features and machine learning techniques. *Internet of Things* 9:100153. doi:10.1016/j.iot.2019.100153.

Kaur, M., G. Kaur, P. K. Sharma, A. Jolfaei, and D. Singh. 2019. Binary cuckoo search metaheuristic-based supercomputing framework for human behavior analysis in smart home. *The Journal of Supercomputing* 76 (4):1–24. doi:10.1007/s11227-019-02998-0.

Khazbak, Y., J. Qiu, T. Tan, and G. Cao. 2020. TargetFinder: A privacy preserving system for locating targets through IoT cameras. *ACM Transactions on Internet of Things* 1 (3):1–23. doi:10.1145/3375878.

Kitchenham, B. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33 (2004):1–26.

Koroniotis, N., N. Moustafa, and E. Sitnikova. 2020. A new network forensic framework based on deep learning for internet of things networks: A particle deep framework. *Future Generation Computer Systems* 110:91–106. doi:10.1016/j.future.2020.03.042.

Lin, P., D.-C. Lyu, F. Chen, S.-S. Wang, and Y. Tsao. 2017. Multi-style learning with denoising autoencoders for acoustic modeling in the internet of things (IoT). *Computer Speech & Language* 46:481–95. doi:10.1016/j.csl.2017.02.001.

Liu, Y., T. Dillon, W. Yu, W. Rahayu, and F. Mostafa. 2020. Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet of Things Journal* 7 (8):6855–67. doi:10.1109/JIOT.2020.2970467.

Mehmood, A., M. Mukherjee, S. H. Ahmed, H. Song, and K. M. Malik. 2018. NBC-MAIDS: Naïve bayesian classification technique in multi-agent system-enriched IDS for securing IoT against DDoS attacks. *The Journal of Supercomputing* 74 (10):5156–70. doi:10.1007/s11227-018-2413-7.

Nesa, N., T. Ghosh, and I. Banerjee. 2018. Non-parametric sequence-based learning approach for outlier detection in IoT. *Future Generation Computer Systems* 82:412–21. doi:10.1016/j.future.2017.11.021.

Rahman, M. A., A. T. Asyhari, L. S. Leong, G. B. Satrya, M. Hai Tao, and M. F. Zolkipli. 2020. Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. *Sustainable Cities and Society* 61: 102324.doi: 10.1016/j.scs.2020.102324.

Rathore, H., C. Fu, A. Mohamed, A. Al-Ali, X. Du, M. Guizani, and Z. Yu. 2020. Multi-layer security scheme for implantable medical devices. *Neural Computing & Applications* 32 (9):4347–60. doi:10.1007/s00521-018-3819-0.

Rathore, S., and J. H. Park. 2018. Semi-supervised learning based distributed attack detection framework for IoT. *Applied Soft Computing* 72:79–89. doi:10.1016/j.asoc.2018.05.049.

Ray, P. P. 2018. A survey on internet of things architectures. *Journal of King Saud University-Computer and Information Sciences* 30 (3):291–319. doi:10.1016/j.jksuci.2016.10.003.

Raza, S., L. Wallgren, and T. Voigt. 2013. SVELTE: Real-time intrusion detection in the internet of things. *Ad Hoc Networks* 11 (8):2661–74. doi:10.1016/j.adhoc.2013.04.014.

RM, S. P., Maddikunta, P. K. R., Parimala, M., Koppu, S., Gadekallu, T. R., Chowdhary, C. L., & Alazab, M. 2020. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Computer Communications*, 160, 139–149.

Santos, L., C. Rabadao, and R. Gonçalves. 2018. Intrusion detection systems in internet of things: A literature review. 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, Spain, IEEE.

Sarker, I. H., and K. Salah. 2019. Appspred: Predicting context-aware smartphone apps using random forest learning. *Internet of Things* 8:100106. doi:10.1016/j.iot.2019.100106.

Sharma, D., I. Mishra, and S. Jain. 2017. A detailed classification of routing attacks against RPL in internet of things. *International Journal of Advance Research Ideas and Innovations in Technology* 3: 692–703.

Shi, Y., X. Zhang, Q. Hu, and H. Cheng. 2020. Data recovery algorithm based on generative adversarial networks in crowd sensing internet of things. *Personal and Ubiquitous Computing* 1–14. doi:10.1007/s00779-020-01428-w.

Sicari, S., A. Rizzardi, L. A. Grieco, and A. Coen-Porisini. 2015. Security, privacy and trust in internet of things: The road ahead. *Computer Networks* 76:146–64. doi:10.1016/j.comnet.2014.11.008.

Skarmeta, A. F., Hernandez-Ramos, J. L., & Moreno, M. V. 2014, March. A decentralized approach for security and privacy challenges in the internet of things. In 2014 IEEE world forum on Internet of Things (WF-IoT) (pp. 67-72). Seoul, Korea (South), IEEE.

Sun, P., J. Li, M. Z. Alam Bhuiyan, L. Wang, and B. Li. 2019. Modeling and clustering attacker activities in IoT through machine learning techniques. *Information Sciences* 479:456–71. doi:10.1016/j.ins.2018.04.065.

Tong, C., X. Yin, S. Wang, and Z. Zheng. 2018. A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data. *Future Generation Computer Systems* 88:552–58. doi:10.1016/j.future.2018.06.023.

Wang, C., S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang. 2019a. Saliencygan: Deep learning semisupervised salient object detection in the fog of iot. *IEEE Transactions on Industrial Informatics* 16 (4):2667–76. doi:10.1109/TII.2019.2945362.

Wang, T., S. Li, and B. Hannaford. 2019b. A model-based recurrent neural network with randomness for efficient control with applications. *IEEE Transactions on Industrial Informatics* 15 (4):2054–63. doi:10.1109/TII.2018.2869588.

Wang, H., Z. Zhao, Z. Wang, G. Xu, and L. Wang. 2019c. Independent component analysis-based baseline drift interference suppression of portable spectrometer for optical electronic nose of internet of things. *IEEE Transactions on Industrial Informatics* 16 (4):2698–706. doi:10.1109/TII.2019.2939645.

Xiao, L., Wan, X., Lu, X., Zhang, Y., & Wu, D. 2018. IoT security techniques based on machine learning: How do IoT devices use AI to enhance security?. IEEE Signal Processing Magazine, 35(5), 41–49.

Xu, R., Y. Cheng, Z. Liu, Y. Xie, and Y. Yang. 2020. Improved long short-term memory based anomaly detection with concept drift adaptive method for supporting IoT services. *Future Generation Computer Systems* 112:228–42. doi:10.1016/j.future.2020.05.035.

Yin, C., Zhang, S., Wang, J., & Xiong, N. N. 2020. Anomaly detection based on convolutional recurrent autoencoder for IoT time series. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(1), 112–122.

Zahra, K., Azam, F., Ilyas, F., Faisal, H., Ambreen, N., & Gondal, N. 2017. Success factors of organizational change in software process improvement: A systematic literature review. Proceedings of the 5th International Conference on Information and Education Technology, Tokyo, Japan.

Zeng, S., X. Tong, N. Sang, and R. Huang. 2013. A study on semi-supervised FCM algorithm. *Knowledge and Information Systems* 35 (3):585–612. doi:10.1007/s10115-012-0521-x.

Zhang, C., and W. Ji. 2020. Edge computing enabled production anomalies detection and energy-efficient production decision approach for discrete manufacturing workshops. *IEEE Access* 8:158197–207. doi:10.1109/ACCESS.2020.3020136.

Zhang, X., X. Zhou, M. Lin, and J. Sun (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6848–56).