Wasit Journal for Pure Sciences **Vol. (2) No. (2)**

IoT Cybersecurity Threats and Detection Mechanisms: A Review

Qassim A. Hadi[1,*] Ali Saeed Alfoudi[1,2] Ahmed Mohsin Mahdi[1,3]

[1]*College of Computer Science and Information Technology, University of Al-Qadisiyah, Al-Qadisiyah, Iraq*

[2]*College of Computer Science, Liverpool John Moores University, Liverpool, UK*

[3]*University of Szeged, collage of Science and Informatics, Hungary*

[*]*Corresponding author's Email: qassim.alzubaidy@qu.edu.iq*

Abstract: *Now a day the Internet of thing (IoT) grab the attention of many researchers and companies due to different direction of utilization. The cyber security of IoT become one of the aspects of the critical challenges. There are many intrusion detection systems (IDSs) to solve different issues of IoT-Cyber security threats. In this article, we review the state-of-the-art of IoT-IDS, focusing on the strategy that was devised and executed, the dataset that was utilized, the findings, and the assessment that was undertaken. Additionally, the surveyed articles undergo critical analysis and statements in order to give a thorough comparative review. Machine learning and deep learning methods, as well as new classification and feature selection methodologies, are studied and researched. Thus far, each technique has proved the capability of constructing very accurate intrusion detection models.*

Keywords: *machine learning – deep learning – IoT networks – anomaly detection*

1. Introduction:

An intrusion Detection System (IDS) is a security mechanism for detecting unauthorized activity, the purpose of which is to prevent intrusion of a system or network[1, 2]. The IDS function detects an attack on the network and issues alerts when such attacks are detected. The decision that is made when an attack is detected is reported to the administrator or is collected using Security Information and Event Management System (SIEM). SIEM integrates output from multiple sources and uses alert filtering techniques to distinguish a malicious attack from false alerts[1]. There are two methods for IDS: the signature-based method and the skew-based method. The signature-based method is based on knowledge-based discovery. The predictions are stored in a prediction database and these predictions are matched against data patterns to detect the attack. Advantage: High detection efficiency for known attacks due to the availability of

anticipation for these attacks. Disadvantage: New attacks cannot be detected due to a lack of predictions in the database. It is a resource-intensive approach because it stores a database of predictions that are compared with data packets for potential interventions. Anomaly-IDS is also called 'behavior-based IDS', any deviation from normal is considered an anomaly. Feature: Discover new and unknown attacks. Disadvantage: high false alarm rate (FAR). It is used to detect unknown malware attacks as new malware is rapidly developed. The machine learning-based method has a better-generalized characteristic compared to the signature-based IDS as these models can be trained according to the applications and hardware configurations[2].

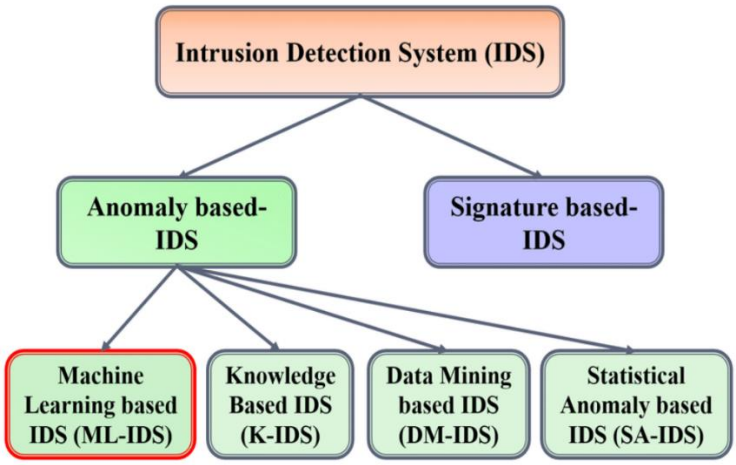

Figure 1: IDS-types

The term "zero-day" only refers to the fact that the developers are not aware of the situation. Once discovered, it was no longer considered a zero-day attack or exploit. Zero-day attacks are targeted exploits designed to take advantage of specific vulnerabilities within widely used software. They are usually packaged as malware and can perform all kinds of malicious chaos: install ransomware, key loggers, worms, spyware, bots, root tools, or a host of threats chained together[3].

The Internet of Things in IDS refers to the network of devices capable of collecting and sharing data with other devices on the same network. Impact of IDS on the Internet of Things an attack on things connected to the Internet is a threat not only to the network or system, but also to servers, applications, and websites, and the movement is partially or completely paralyzed. The Internet of Things is not a single technology, but a mixture of different hardware and software technologies. The Internet of Things provides solutions based on the integration of information technology, which refers to the hardware and software used to save, retrieve and process data, and communication technology that includes electronic systems used to communicate between individuals or groups[4].

The main characteristic of the Internet of Things:

interconnection: anything can be connected to the global information and communication infrastructure.

Services related to things: The Internet of Things allows the provision of services related to things within the constraints of things, such as protection of privacy and semantic consistency between physical objects and virtual objects connected to them.

Heterogeneity: Devices in IoT are heterogeneous depending on different device platforms and networks. They can interact with other devices on different networks.

Dynamic Changes: The device's state changes dynamically e.g. sleep and wake, connected and/or disconnected as well as device context including location and speed.

Huge range: The number of managed devices and this connection to each other will be at least an order of magnitude larger than devices connected to the current Internet.

Security: As we gain the advantages of the Internet of Things, we must not forget about security. We must design for safety. This includes the data, the integrity of the physical well endpoints, the networks, and the data that is transmitted through it all means creating a security model that will scale.

Connectivity: Connectivity enables compatibility with network access. Access is obtained on a network while compatibility provides the combined ability to consume and produce data[5].

There are many different datasets: BoT-IoT Dataset This dataset is rich in features and types of attacks. This study aims to analyze contributions toward the imbalance of the data set. Bot-IoT includes both normal IoT-related network traffic and other network traffic, along with various types of attack traffic commonly used by bot networks. The full data set contains about 73 million instances (big data). Bot-IoT-trained models can detect various bot attacks in the Internet of Things (IoT) networks[6].

N-BaIoT dataset This dataset consists of 115 real number attributes and many instances: 7062606. It suggests real traffic data, collected from 9 commercial IoT devices originally infected by Mirai and BASHLITE[7]. DS2OS traffic tracks this dataset and description: A virtual IoT environment is created using a distributed intelligent space distribution system (DS2OS) that contains a set of IoT-based services such as temperature controller, window controller, lighting controller, and so on. The user and services are captured and stored in a CSV file format. In the data set, there are 357,952 samples and 13 features. The data set contains 347,935 normal data and 10,017 anomalous data and contains eight categories that were classified. The eight attack categories are Denial of Service (DoS), data type checking, malicious control, malicious process, scanning, spyware, misconfiguration, and normal. The dataset is free to use and available on the Kaggle website[8].

The layout of the paper is organized as follows: section 3 discusses the Proposed model steps. section 3 discusses the types of attacks. Section 4 discusses IDS in IoT Studies. Section 5 discusses the Analysis of IDS in IoT studies. finally, section 5 discusses the challenges of IDS in IoT.

2. Attack types of IoT:

- **Spoofing attack**: It refers to the use of credentials belonging to others to access an inaccessible service. Credentials can be obtained directly from a device or installed on the communication or phishing channel. There are three types of spoofing i) IP address spoofing; ii) ARP spoofing; and iii) DNS server spoofing[9].

- **routing attacks**: The most direct attack against a routing protocol is to target the routing information exchanged between nodes. By intimidating, altering, or replaying routing information, adversaries may be able to create routing loops, attract or block network traffic, extend or shorten source routes, generate false error messages, split the network, increase overall latency, etc.[10].

- **Sinkhole attack**: this attack is the most threatening attack on the network layer that sends fake information assuming it is the shortest path to the base station so that the entire Traffic network is drawn towards it. Present an imaginary path as the optimal routing path[11].

- **forwarding attack**: It is possible to launch DOS attacks that target malicious nodes selectively. This attack is primarily aimed at disrupting routing paths; However, it can be used to filter any protocol. For example, an attacker could redirect all RPL control messages and drop the rest of the traffic. This attack has serious consequences when combined with other attacks, for example, pelvic attacks[12].

- **black hole attack**: One or more malicious nodes advertise themselves as the best ways to (partially or completely) drop data packets that are routed through them, to disrupt the normal network traffic[13].

- **wormhole attack**: At least two malicious nodes communicate using a separate wired or wireless link called a "tunnel" to forward packets faster than normal paths[13].

- **tampering attack**: It is classified as i) tampering with the device, and 2) tampering with data. Device tampering can be carried out easily especially when the IoT device spends most of the time unattended. It can be easily stolen without being noticed and used maliciously. The device can be Stolen as hardware or as software. Data tampering involves malicious modification of data for example data stored in databases or data transmission between two devices[5].

- **Repudiation attack**: By passing controls to properly track and record users' behavior, an application or system is vulnerable to disavowal attacks. A malicious user can use this technology to change the authorship information of their actions, which leads to the recording of inaccurate data. Similar to spoofing emails, it can be used to process data on behalf of others[5].

- **information disclosure**: It is the act of disclosing information to an entity that does not have permission to see it. This includes accidental exposure, targeted attack, and inference or association. An attacker can obtain information by eavesdropping on the network channel, physically gaining access to the device, or by accessing the device over the network[9].

- **elevation of privilege**: It is when an unblocked user gets privileged access to a device/service. This can be achieved by installing a fraudster in the system pretending to be another device, having privileged access to the system[6].

- **MITM (Man-In-The-Middle)**: A type of attack where a malicious third party secretly controls the communication channel between one or more endpoints. A MITM attacker can intercept, alter or replace

the communications traffic of the targeted victims (this distinguishes MITM from simple eavesdropping). Moreover, the victims are unaware of the intruder, which is why the communication channel is believed to be protected. The attack can be carried out in different communication channels such as GSM, UMTS, Long Term Evolution (LTE), Bluetooth, Near Field Communication (NFC), and Wi-Fi. The targets of the attack are not only the actual data flowing between the endpoints but also the confidentiality and integrity of the data itself[14].

- **cloning Nodes**: These types of attacks are known as identity attacks. In a clone identifier attack, the attacker copies a valid node identity to multiple physical nodes; However, the attacker copies many logical identities onto a single physical node in a Sybil attack. Such an attack enables a malicious user to take control of the system, insert false information, disable functions, etc.[8].

- **Denial-of-Service(DoS)**: The most common attacks especially in IoT networks/fog related to social IoT such as smart cities, etc. Indicates a property that is inaccessible when requested by an authorized user. The system must have the ability to continue running even when some unwanted actions are performed by malicious users. This class of attacks can be carried out by stealing the device, manipulating its software, or disrupting the communication channel[9].

- **Distributed Denial of Service(DDoS)**: this attack is performed by multiple vulnerable nodes together from different geographic locations. Furthermore, a DOS attack involves a malicious attacker attempting to consume network resources, targeting the CPU time and/or bandwidth of legitimate users by flooding the system with rogue and amplifying traffic. To conduct an effective DDoS attack, bots are used. They are networks of devices infected with the Internet[15].

3- **IDS in IoT studies**

Internet of Things (IoT) – Intrusion Detection System(IDS) focused on the efficiency of the NIDS in IoT networks. Alaa Alhowaide et al.[16]] utilized a classification model which was automatically selected and build ensemble detection models based on machine learning. The evaluated model is Based on F_scores, ROC-AUC scores, and accuracy. The proposed model achieved 0.99, 0.95, 1, and 0.99 F scores and 1, 0.98, 1, and 1 ROC-AUC scores when applied to the NSL-KDD, UNSW-NB15, BoTNe-TIoT, and BoTIoT benchmark datasets. This research is incapable of detecting the attack type. moreover, the proposed MSM algorithm incorporates different decision combination methods and more efficient measurements. However, the proposed ENCLF models were tested on session–based datasets. In the future, the challenge is how to build a model that is able to detect the type of attack, not just detect if there was an attack or not.

Shafiq et al.[17] had focused on the high dimensionality of the IoT network data. The authors proposed the corrAUC model as a wrapper feature selection technique to select highly relevant features. The proposed model combined Correlation Attribute Evaluation (CAE) with the Area Under Roc Curve (AUC) metric to overcome the problem of effective feature selection by using a specific machine learning (ML) algorithm. the integrated TOPSIS and Shannon Entropy based on a bijective soft set were utilized as fitness functions. The accuracy, precision, and sensitivity were utilized as evaluation metrics when the bot-IoT dataset was utilized as a benchmark dataset. The proposed mode achieved an efficient best result up to >96%.

YANPING SHEN et al.[18] focused on effectively producing individual learners with a strong ability for generalization and large differences in ensemble-based IDS efficiency. The authors propose an Ensemble pruning framework. It is the intermediate phase between the construction of the sub-classifiers and the final decision; the main job of ensemble pruning is to reduce the size of the classifiers for the ensemble. Moreover, the proposed framework is based on the selection using the bat algorithm (BA) to choose the learner subset for intrusion detection in this paper. Moreover, the ELM, generated based on random subspace, is selected as the core learning algorithm in the ensemble. The selected models are combined using the majority voting method. KDD99, NSL, and Kyoto datasets are used to evaluate and know the results of the method. The model utilizes old datasets KDD Cup'99, NSL-KDD, and Kyoto for evaluation of the proposed model. Further, the model realizes lower detection accuracy for the U2R attack.

Daniele Midi et al.[19] focused on the performance efficiency of Knowledge-driven Adaptable Intrusion Detection for the Internet of Things. The authors proposed Kalis, a self-adapting, knowledge-driven expert Intrusion Detection System, where capable of detecting attacks in real-time. the Kalis is an overall approach that detects attacks for IoT that do not target individual protocols or applications and adjust strategy for specifying network characteristics. The evaluation method shows, that Kalis is effective and efficient in detecting attacks in IoT systems. the model depends on accuracy, CPU usage, RAM usage, and detection rate to evaluate the results. The proposed model achieved 100% accuracy, 91% detection rate, 0.19% CPU usage and 13978.62% RAM usage (kb). This method may not be appropriate for limited computing objects. kalis propose gathering time Publishing that may not important for resource-limited sensors that may be resource-limited in comparison to WSN nodes.

Yakub Kayode Saheed.[20] focused on increasing the performance of the IDS in IoT by reducing the data dimensionality. The authors proposed the PCA algorithm for dimensionality decreasing to a select few components. The classifier XGBoost, CatBoost, K Nearest neighbor (KNN), Support vector Machine(SVM), and Quadratic discriminant analysis(QDA)to classify the intrusion detection data. The proposed method evaluated from where of validation data-set, accuracy, the area under the curve, recall, F1, precision, kappa, and Mathew correlation coefficient (MCC). The dataset utilized in this paper UNSW-NB15 dataset. The best results from this work are an accuracy of 99.9% and an MCC of 99.97%. This paper used the old dataset UNSW-NB15 and used a single model rather than an ensemble model.

Amar Amouri et al.[21] focused on Enhancing the performance of the IDS due to the distributed nature and the limited resources available in the IoT networks. The authors proposed a new model that is composed of two stages; stage one collects data through dedicated sniffers (DSS) and generates the CCI which is sent in a periodic fashion to the super node(SN), and in stage two the SN performs the linear regression process for the collected CCIs from different DSs in order to differentiate the benign from the malicious nodes. The evaluated model depends on the power level, node velocity, F1 score, false positive rate (FPR), and true positive rate (TPR). The proposed model achieved the best results up to, 98% for high power/node velocity scenarios On the other side they drop to around 90% for low power/node velocity scenarios, and the F1 score varied between 93% and 99.36%. This work finds the false positive rate (FPR) that ranged between 1.3% and 12% across various scenarios this is a restriction on IDS.

Ruhul Amin et al.[22] Focused on the large amount of IoT network data that composed form the different smart devices in IoT. The authors designed new architecture for a distributed cloud environment where the private cloud stores confidential information using the Internet of Things (IoT) technique. To get secure access to confidential information from any private cloud server of the distributed system, this article designs a standard authentication protocol that resists all kinds of security attacks and provides important features such as user anonymity. Mutual authentication proof has been done using BAN logic and the protocol simulation using AVSIPA results ensures the security and safety of the protocol. Moreover, the informal cryptanalysis of the proposed protocol ensures that the protocol is security attacks protected under the hardness assumption of the hash function. This paper shows the security vulnerabilities in Cloud Computing (CC), but the protocol is weak to face Password Guessing Attacks, secret guidance, and Users' inability to track.

Prosanta Gope et al.[23] focused on Improved RFID authentication schemes in IDS in IoT. The proposed model is utilizing an RFID-depended authentication structure for distributed IoT (Internet of Things). RFID uses electromagnetic fields to automatically define and track tags attached to objects. The model is evaluated in terms of Mutual Authentication, which provides strong anonymity, saving, Forward Security, scalability, and Security resettlement. in this paper, one of the main problems is The back server is very powerful so that the server can know all communication RFID tags, and when the server hack, the attacker can get all secret data. RFID schema suffers from physical and cloning attacks so It is a real concern.

Marc Barcelo et al.[24] had focused on planning to formulate service distribution problem (SDP) in IoT Cloud networks mathematically and focus on Energy consumption as a major driver of current cloud operating costs and distinguish the heterogeneous pool of resources of IoT-Cloud network. The authors present a network-flow-depend linear programming solution that optimizes the distribution of cloud services with random function relationships (e.g., service chaining) over a distributed cloud network. However, the proposed work does not take into account the increased flexibility that creates when presenting the access network and the device layer into the virtualized infrastructure, aspects that are important for the efficient delivery of IoT services. the evaluation method in this work depends on computing, sensing, transport, capacity, and energy efficiency. The proposed model achieved the best results up to 80% generally to reduce energy consumption While ensuring more robust latency restrictions from one side to the other.

John OcheOnah et al.[25] Most of the time, the fog's nodes produce massive amounts of data because of the direct contact of the end-users and the lack of available computer resources. The use of fog machines might lead to security problems. Due to the inefficiency of traditional IDS, implementing them directly on a fog computing platform may be inappropriate. Fog computing requires the use of Efficient IDSs that can deal with massive databases. This article proposed a Genetic Algorithm Wrapper-Based Feature Selection and Nave Bayes for Anomaly Detection Model (GANBADM) in a Fog Environment that eliminates superfluous attributes to minimize time complexity with high accuracy. GA is used as a random search technique with Naïve Bayes classifier as a classification method. The evaluation metrics are accuracy, precision, F1 score, and execution time as performance metrics. The result is a 98% overall true positive rate, 0.6% as False Positive Rate, and a 99.73% accuracy when utilizing NSL-KDD Dataset.

Nour Moustafa et al.[26] focused on Designing a model to improve the protection of the IoT network. The authors present a NIDS based on an AdaBoost ensemble learning algorithm that takes statistical flow features as input for recognizing malicious botnet activities. Moreover, the AdaBoost ensemble learning methodology is used to combine three classification techniques of DT, NB, and ANN for detecting and improving the performance of NIDS. The correlation coefficient is utilized for selecting the lowest correlated features that have the potential characteristics of legitimate and malicious patterns. The evaluation model is based on Accuracy, Detection Rate (DR), False Positive Rate (FPR), and ROC curves in evaluating the performance of the model. Accuracy is 99.54%, DR is 99.86%, FPR is 0.01% and ROC curves when utilized UNSW-NB15 dataset. In fact, this study is robust to overfitting, performs better than a single classifier and It reduces variance but it Increased time complexity, due to the use of multiple classifiers in parallel.

Liqun Liu et al.[27] in this study focus on the optimization of efficiency and effectiveness of intrusion detections. The authors of this research proposed an objective prejudgment-based intrusion detection, and a frequency self-adjustment algorithm for IoT was proposed. In this algorithm, the huge data flow is integrated and analyzed. More specifically, the data is classified using the clustering algorithm: this research uses PCA for reducing data dimensionality and eliminating features with low discriminations. And Suppressed fuzzy clustering (SFC) algorithm to clustering reduced as high-risk and low-risk data clusters. detection duration (T), accuracy (P), and false alarm rate (F) were employed as evaluation parameters. Detection Duration is in the 40s, Accuracy is 97.1%, and the False alarm rate is1.5% shown in this work. In this study, the efficiency was promising. Nevertheless, it will be inefficient if the data volume increases.

The main objective of this work is to build machine learning models to identify attacks in IT networks. K. V. V. N. L Sai Kirana et al.[28] employ Machine Learning classifiers; SVM, Adaboost, decision trees, and Naïve Bayes to classify data into normal and attack classes. In their work, they used Node MCU-ESP8266, DHT11-sensor, and a wireless router to simulate an IoT environment. They then built an adversary scheme with a computer, which implements poisoning and sniffing attacks on the IoT environment. The steps they followed while building their system are as follows: Develop a testbed to mimic an IoT-based environment develop an attack-like system to obtain attack data Obtain the flow of data in the system and generate normal and attack scenarios feature Build Machine Learning and DL methods to identify and categorize network attacks. The evaluation model depends on accuracy, error rate, sensitivity (recall), specificity, precision, F1, detection rate, and false alarm rate measures. The dataset is Data Collection from think Speak.

Abhishek Verma and Virender Ranga [29] had focused is basically on utilizing ML classification algorithms for building IDS in order to secure IoT against DoS attacks. The authors are based on seven ML algorithms random forests, ad boost, gradient boosted machine, extremely randomized trees, classification and regression trees, and multi-layer perceptron to evaluate the performance of the proposed model. The evaluation model is Based on the accuracy, specificity, sensitivity, and false positives. the proposed RF-based IDS outperforms the ensemble of Random tree + Naive Bayes and single classifiers like NB Tree and Multilayer perceptron. statistical analysis based on Friedman's ranking showed that the ensemble of 800 trees achieves the best results when utilizing CIDDS-001, UNSWNB15, and NSL-KDD datasets.

Mengmeng Ge et al.[30] focused to improve both the false positive and the false negative of the IDS detection in IoT. The authors proposed multiclass and binary class schema by utilizing feed-forward neural networks(FNN). The evaluation model depends on accuracy, precision, recall, and F1 score to evaluate the result and know the effect of the model. This model achieved 99.414% accuracy when using binary class and 82% accuracy when using multiclass class. The proposed model in this work utilized the Bot-IoT dataset. In this work, the multiclass suffers from uncertainty in results due to the field information for the individual packet could not capture certain attack behavior on a large scale so the binary class shows the best results in this field.

Yazan Otoum et al.[20]] Focused on Implementing an efficient intrusion detection system (IDS) in the Internet of Things (IoT) by defining data as normal or severe anomalies in various attacks such as (DoS, U2R, R2L, and probe). The authors combined the spider monkey optimization (SMO) algorithm and the stacked-deep polynomial network (SDPN) to implement new IDS. The SMO was utilized to reduce the network data dimensionality by selecting high relevant feature subset. Moreover, the SDPN was utilized for detecting the attack behavior. The evaluation model is based on accuracy, precision, recall, and F1-score to evaluate the proposed model. The proposed model achieved the best result in terms of accuracy (99.02%), precision (99.38%), recall (98.91%), and F1 score (99.14%) when utilizing the NSL-KDD dataset. In fact, the size of the dataset caused complexity when using deep learning algorithms.

Zhihong Tian[31] Focused on designing IDS in IoT for detecting URL attacks. The proposed model utilized deep learning techniques to detect attacks from URLs by using Natural Language Processing (NLP) and Convolutional Neural Networks (CNNs). The CNN was utilized in Feature Discriminator. The CBOW, one of the word2vec models of NLP, was utilized to represent words in normalized URLs with vectors. The evaluated model utilized accuracy, recall, FP, and precision to evaluate the result and determine its effectiveness. The proposed method achieved 99.410% accuracy, 98.91% in TPR and 99.55% in DRN demonstrate when utilizing three datasets HTTP Dataset CSIC 2010, FWAF, and Http Params Dataset. In this work, the decisions are limited and not comprehensive decisions and do not use all deep learning techniques in optimization so the optimization ratio is different.

Muhammad Almas Khan et al.[32] Focused on the protection of Message Queuing Telemetry Transport (MQTT) in the Internet of Things (IoT). The authors utilize a Deep Neural Network (DNN) to detect the intrusion in the Message Queuing Telemetry Transport (MQTT) protocol that is widely used publish–subscribe-based to deliver sensor or event data. the evaluation model utilized Uni-flow, Bi-flow, and Packet-flow to evaluate the results. The proposed method achieved the best results of 99.92%, 99.75%, and 94.94% accuracies for Uni-flow, Bi-flow, and Packet-flow, respectively. The first dataset utilized includes MQTT-IoT IDS 2020 and another dataset with three various types of attacks, such as Man in the Middle (MitM), Intrusion in the network, and Denial of Services (DoS).

Mohamed Amine Ferrag et al.[33] Focused on the DDoS attack in Agriculture 4.0 of IoT network. The authors introduced deep neural networks including convolutional neural networks(CNN), deep neural networks(DNN), and recurrent neural networks (RNN). The evaluation model depends on detection rate (DR), false alarm rate (FAR), precision, F-score, recall, TNR, FAR, ROC Curve, and accuracy. Each model's performance is studied within two classification types (binary and multiclass), and the result

achieved an accuracy of 99.95% for binary traffic detection and 99.92% for multiclass traffic detection. The proposed datasets are the CICDDoS2019 dataset and the TON IoT dataset which contain various types of DDoS attacks.

Manuel Lopez-Martin et al.[34] focused on improving IDS that deal with unbalanced, noisy, and large network data which have fast prediction and training times. The authors utilized shallow linear models, which used kernel approximation (KA) theory. The proposed model depends on Neural networks (NN) with linear activations. the evaluation model depends on comparing with machine learning techniques such as (MLP, CNN, BM AdaBoost, SVM, and RF Linear model) in terms of accuracy, F-1, recall, and precision. The proposed model gives the best results when utilizing the Moore dataset which achieved 99.8% accuracy when utilizing NSL-KDD, UNSWNB15, and Moore datasets.

Pushparaj Nimbalkar and Deepak Kshirsagar[35] Focused on reducing the noisy network traffic in IoT networks. The authors proposed a new model that utilized Information Gain (IG) and Gain Ratio (GR to select features, then obtains feature subsets using insertion and union operations on subsets obtained by the ranked top 50% IG and GR Features, finally using JRip classifier to measure the performance of the model. The evaluation model utilized accuracy (ACC), detection rate (DR), model built-up time (B. Time), and false alarm rate (FAR) to evaluate the result. The model achieved higher accuracy and detection rate of 99.9993%, and 99.5798% respectively, with JRip using 16 features on the BoT-IoT dataset, and The KDD Cup 1999 dataset's system validation also optimized accuracy and detection rate of 99.9920% and 99.9943% to detect DoS attack using 19 features with JRip.

Mrs.G.Parimala and Dr.R.Kayalvizhi[36] focused on feature selection and reduced dataset volume. The authors proposed a hybrid model that combined (SMO) and (CRF). The Conditional Random Field (CRF) and spider monkey optimization (SMO) were utilized to define the features useful in the dataset. The Convolutional Neural Network(CNN) was utilized to classify the dataset as normal and the attacks. The evaluation model is based on detection accuracy, time, and false positive rate to evaluate the method and know the performance. The public NSL KDD dataset consists of 41 features but after applying the model the dataset becomes 38 features.

4-Analysis of IDS in IoT studies:

In the table below (Table 1), we review the summary of previous studies and the most important operations that were performed:

references	Proposed solution	Dataset	Evaluation model	disadvantage
Alaa Alhowaide et al. [17]	an ensemble classification model	NSL-KDD UNSW-NB15 BoTNeTIoT, BoTIoT	-F- scores -ROC-AUC scores.	incapable of detecting the attack type
Muhammad Shafiq[18]	CorrAUC proposed model	Bot-IoT dataset.	accuracy, precision, sensitivity, and specificity metrics	The results have not been clarified zero-day attack

YANPING SHEN et al. [19]	extreme learning machine (ELM)	KDD99, NSL-KDD Kyoto datasets	accuracy and robustness	The model utilizes old datasets and doesn't use for IoT
Daniele Midi et al [20].	Kalis	No dataset	accuracy, CPU usage , RAM usage, and detection rate	This method may not be appropriate for limited computing objects.
Yakub Kayode Saheed [21].	Min-max normalization, PCA algorithm, XGBoost, CatBoost, KNN, SVM, and QDA	UNSW-NB15	accuracy, the area under the curve, recall, F1, precision, kappa, and Mathew correlation coefficient (MCC).	The model utilizes old datasets and does not use for IoT
Amar Amouri et al [22].	Random Way Point (RWP), and Gauss Markov (GM).	No dataset	power level, node velocity, F1 score, false positive rate (FPR), and true positive rate (TPR).	false positive rate (FPR) that ranged between 1.3% and 12%
Ruhul Amin et al [23].	Cloud Computing (CC)	No dataset	Burrows-Abadi-Needham(BAN) and AVISPA tool	shows the security vulnerabilities in Cloud Computing (CC)
Prosanta Gope et al[24].	RFID	No dataset	Mutual Authentication, Provides strong anonymity, saving, Forward Security, scale, ability, and Security resettlement	- RFID suffers from physical and cloning attacks so It is a real concern. - The back server is very powerful so that the server can know all communication RFID tags.
John Oche Onah et al [26]	GA with Naïve Bayes	NSL-KDD	accuracy, precision, F1 score, and execution time	The model utilizes old datasets and doesn't use for IoT

Nour Moustafa et al [27]	AdaBoost ensemble learning	UNSW-NB15	Accuracy, Detection Rate (DR), False Positive Rate (FPR), and ROC curves	Increased time complexity due to the use of multiple classifiers in parallel.
Liqun Liu et al [28]	- PCA algorithm - Suppressed fuzzy clustering (SFC) algorithm	No dataset	detection duration (T), accuracy (P), and false alarm rate (F)	inefficient if the data volume increases
K. V. V. N. L Sai Kirana et al [29]	- SVM with NPSO, - Adaboost with DT	Data Collection from Think Speak	accuracy, error rate, sensitivity (recall), specificity, precision , F1-score detection rate	The results have not been clarified of the zero-day attack.
Abhishek Verma et al [30]	- random forests - ad boost - gradient - boosted machine. - extremely randomized trees, - classification and regression trees, - multi-layer perceptron	-CIDDS-001 - UNSWNB15 - NSL-KDD	-accuracy -specificity -sensitivity - false positive	The data is old and does not match the work environment
Mengmeng Ge et al [20].	feed-forward neural networks(FNN).	Bot-IoT	-accuracy, -precision, -recall -F1 score	the multiclass suffers from uncertainty in results
Yazan Otoum et al [31]	- spider monkey optimization (SMO) - stacked-deep polynomial network (SDPN)	NSL-KDD	-accuracy, -precision -recall -F1-score	complexity when using deep learning algorithms
Zhihong Tian et al[32]	- Natural Language Processing (NLP) - Convolutional Neural Networks (CNNs)	- HTTP Dataset CSIC 2010 - FWAF - HttpParams	-Accuracy -recall -FP -precision	the decisions are limited and not comprehensive decisions and do not use all deep learning techniques in optimization so

				the optimization ratio is different.
Muhammad Almas Khan et al [33].	- Deep Neural Network (DNN)	MQTT-IoT IDS 2020	-Uni-flow - Bi-flow -Packet-flow	Increasing the complexity of the model, makes it consume time and resources during implementation, and this is not in line with IoT
Mohamed Amine Ferrag et al[34].	-convolutional neural networks(CNN) - Deep Neural Network (DNN) - recurrent neural networks (RNN).	- CICDDoS2019 - TON IoT	-detection rate (DR) -false alarm rate (FAR) - precision, -F-score -recall -TNR -FAR -ROC Curve -accuracy	complexity when using deep learning algorithms in the model
Manuel Lopez-Martin et al [35]	kernel approximation (KA) theory	- NSL-KDD - UNSWNB15 - Moore datasets	-Accuracy -F-1 -recall -precision.	The dataset doesn't use for IoT
Pushparaj Nimbalkar et al [36]	- Information Gain (IG) - Gain Ratio (GR)	- Bot-IoT - KDD Cup 1999	-accuracy (ACC) -detection rate (DR) -model built-up -time (B. Time) -false alarm rate (FAR)	Low results for attack classes comp red to classes normal
Mrs.G.Parimala et al [37]	- spider monkey optimization (SMO). - Conditional Random Field (CRF)	NSL-KDD	-detection accuracy time, - false positive rate	Low performance in a real-world environment because the dataset is old.

Table 1: previous studies

5- The different challenges of IoT anomaly-based intrusion detection systems

In the Internet of Things environment, there are many challenges facing the researchers due to the verity nature of IoT data such as the complex, data imbalance and redundancy. Therefore, we review some of these challenges according to the device capability and the data generation of IoT environment:

1. IoT devices limitation:

Generally, IoT devices have limited capacity due to device limitations in terms of memory capacity, processor, and battery lifetime. In the following section, we discuss the IoT device challenges according to different environments and computational technologies.

Heterogeneity:

The environment of the Internet of Things is a wide environment where many devices, protocols, and different standards are connected to it, and therefore the heterogeneity between these devices is very large. Therefore, the data is varied and poses a real challenge for the researcher. These devices and protocols, and Because these devices and protocols are not homogeneous, they give a variety of data[5]. One of the most important problems of heterogeneity is that it brings harmful data or low-quality data and thus affects the functioning of the system. Data heterogeneity can be classified in terms of (data quality, data quantity, data quality, etc.), where the devices are different, and therefore the heterogeneity between devices leads to the emergence of gaps, and thus makes it easier for attacks to exploit these gaps and attack the system, and this would affect the performance of the model and thus reduce its efficiency at work. One of the suggested solutions to identify heterogeneity is to implement a smart central server, which relies on reinforcement learning, thus achieving better performance[37]. The Internet of Things systems are widely distributed systems, and therefore it is difficult to deal with them, and therefore the challenge appears to us in terms of protecting these systems from penetration, due to the heterogeneity that exists between them.

Time complexity and memory usage:

Time complexity can be defined as the computational complexity that describes the amount of time it takes for the model to find results, which is directly proportional to the size of the data, as an increase in the size of the data leads to an increase in the time complexity. The data of the Internet of Things is flowing data, and therefore it is large data, and therefore there will be time complexity. This is due to the large data volume[38]. One of the reasons for the complexity of time is also the diversity of data, as the heterogeneity of the devices leads to a variety of data, and thus it becomes difficult to process and takes more time for processing. To reduce the complexity of time, you must use the methods and techniques that are used to get rid of unnecessary and duplicate data that consumes a long time in processing and is useless because it may be harmful data[39].

Data in the Internet of Things requires a very large amount of memory, for several reasons, the most important of which are (data diversity, data size, and data speed) and thus pose a challenge to researchers. Due to the pressure on the Internet infrastructure due to the volume of data, one of the proposed solutions is cloud computing, which would solve the problem of storage and processing of data and thus reduce the memory problem for the Internet of Things and thus increase the efficiency and performance of the system and also allow access to data remotely and thus avoid any delay[40].

Optimal Data Capture and Processing:

A major issue is created within the framework of the Internet of Things with more information transmitted on the system. Because a huge amount of the information is meaningless to the client, the methods of filtering the information will be optimally before storage and will rise as an important search area. Collecting data from devices, shaping the topology, forwarding packets, optimizing resources and power, optimizing coverage, efficient assignment of tasks, and security are important challenges in the IoT environment[41]. The process of collecting and processing data in a short time and with good results is one of the problems in the Internet of Things environment, as traditional networks take a long time to achieve a satisfactory data delivery rate[42]. The data in the environment of the intensification of things is very large and flowing data, and therefore it is difficult to process and deal with it, because of the continuous change in the shape and size of the data. Therefore, it is a great challenge in the process of collecting and processing data. And because the shape of the data is diverse and different, we need advanced technologies to link this data with each other. Some of them are configured to be processed and extract the necessary data from them[43].

Interoperability:

The concept of interoperability can be defined as the ability to create systems or devices that cooperate with each other in an efficient manner. The basic idea of the proposed architecture is to divide the IoT environment into small spaces to facilitate its management[44]. The semantic information broker uses SIB to provide a way for agents to share semantic information with each other, and also provides real-time monitoring and updating of the physical world. The main note of the architecture is the performance after using the proxy interaction operations scale very well and it also allows interacting with the physical world in real time. The architecture needs tools to support the development and deployment of devices and applications in future IoT systems[45]. ISO/IEC defines interoperability as “the ability to communicate, execute programs, or transfer data between different functional units in a way that requires the user to have little or no knowledge of the unique properties of those units[46]. In a broader perspective, interoperability is defined by the IEEE It is defined as “the ability of two or more systems or components to exchange information and use the information that has been exchanged. According to this definition, interoperability is achieved by setting standards. Interoperability in the Internet of Things can be defined as the ability of two systems to communicate and share services with each other[47].

2. Different challenges in IoT-IDS datasets within machine learning:

The machine learning algorithms are used widely to solve the problems of the IoT-IDS, where The results showed the accuracy of the machine learning algorithms in such challenges. We review the following challenges:

Imbalanced Data:

Imbalanced classification refers to a classification predictive modeling problem where the number of examples in the training dataset for each class label is not balanced. That is, where the class distribution is not equal or close to equal and is instead biased or skewed. The unbalanced classification problem is an example of a classification problem in which the distribution of examples across known categories is biased or skewed. Distribution can vary from slight bias to severe imbalance where there is a single example in the minority stratum to hundreds, thousands, or millions of examples in the majority stratum or strata[48]. The machine learning algorithms utilized in this challenge are over-sampling, under-sampling, and smooth. They are utilized in data mining and data analytics to modify unequal data classes to create balanced data sets. These data analysis techniques are often used to be more representative of real-world data. The serious limitation of the sampling methods is that it involves biased selection and thereby lead us to draw erroneous conclusions. Bias arises when the method of selection of the sample employed is faulty. Relative small samples properly selected may be much more reliable than large samples poorly selected[49].

Missing values:

missing values is a common and unavoidable challenge in the data processing and analysis phase, the reason for this is due to failure to collect the samples correctly, or not to store the data, the presence of restrictions in the data acquisition process, and thus the loss of this data occurs, and thus it has a noticeable impact[50]. One of the results of the missing values is the poor knowledge extraction process, as well as the wrong conclusion process, and thus affect the work of the system, as well as the loss of efficiency and accuracy in the model extraction process. The messing value in term numeric utilized mean technique and the missing value in term nominal utilized most frequent. The strategy is “mean”, which replaces missing values with the median value of the column. The “most frequent” (which replaces missing values with the most common value in the column) and “constant” (which replaces missing values with a constant value). The weakness of mean technique is “it reduces the variance”. In most frequent must sure don't have very skewed class distributions[51].

Data redundancy:

The term big data refers to data that includes increasing volumes, variety, and flow velocity, and it can be referred to by the term (3V). When the data is large in terms of sampling and prediction, the algorithms face a big problem, and therefore it is difficult to deal with it[52]. This problem is solved by selecting only effective data using processing techniques, and one of the most important of these techniques is the feature selection process, as it is one of the most important operations in the pre-processing stage[53]. The feature selection process can be defined as the process of selecting relevant and influential features from the raw data set to reduce unnecessary features. The process of selecting features reduces the search area that is determined by the features, and thus the learning process is easy and simple, and also reduces memory consumption. The selection of features can be used in the data collection process, thus reducing the time and also taking the necessary samples in the early stages[54]. The set of features that have been chosen is a subset of the original data set, as it describes the original data appropriately and thus facilitates the process of understanding and working on it.

Conclusion

In conclusion, this article reviewed 22 effective techniques leveraging various machine learning and optimization processes in intrusion detection systems. The suggested analysis in this survey focused on accuracy as the primary criterion. We also checked for processing time and were disappointed to discover that they lacked any system performance data, including processing time. By and large, all groups demonstrated superior skills in terms of the accuracy measure. Additionally, we examined how machine learning methods may be used for cybersecurity and other security-related challenges. In terms of the present research, conventional security solutions have garnered considerable attention, whereas security systems based on machine learning techniques have received less attention. We've reviewed pertinent security research for each widely used technique. This article will offer an overview of the conceptualization, understanding, modeling, and reasoning processes involved in cybersecurity data science.

References

1. Smaha, S.E. *Haystack: An intrusion detection system*. in *Fourth Aerospace Computer Security Applications Conference*. 1988. Orlando, FL, USA.
2. Yassin, W., et al. *Signature-Based Anomaly intrusion detection using Integrated data mining classifiers*. in *2014 International symposium on biometrics and security technologies (ISBAST)*. 2014. IEEE.
3. Blaise, A., et al., *Detection of zero-day attacks: An unsupervised port-based approach.* Computer Networks, 2020. **180**: p. 107391.
4. Patel, K.K., S.M. Patel, and P. Scholar, *Internet of things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges.* International journal of engineering science and computing, 2016. **6**(5).
5. Chaabouni, N., et al., *Network intrusion detection for IoT security based on learning techniques.* IEEE Communications Surveys & Tutorials, 2019. **21**(3): p. 2671-2701.
6. Koroniotis, N., et al., *Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset.* Future Generation Computer Systems, 2019. **100**: p. 779-796.
7. Abbasi, F., M. Naderan, and S.E. Alavi. *Anomaly detection in Internet of Things using feature selection and classification based on Logistic Regression and Artificial Neural Network on N-BaIoT dataset*. in *2021 5th International Conference on Internet of Things and Applications (IoT)*. 2021. IEEE.
8. Pahl, M.-O. and F.-X. Aubet. *All eyes on you: Distributed Multi-Dimensional IoT microservice anomaly detection*. in *2018 14th International Conference on Network and Service Management (CNSM)*. 2018. IEEE.
9. Atamli, A.W. and A. Martin. *Threat-based security analysis for the internet of things*. in *2014 International Workshop on Secure Internet of Things*. 2014. IEEE.
10. Karlof, C. and D. Wagner, *Secure routing in wireless sensor networks: Attacks and countermeasures.* Ad hoc networks, 2003. **1**(2-3): p. 293-315.

11. Bostani, H. and M. Sheikhan, *Hybrid of anomaly-based and specification-based IDS for Internet of Things using unsupervised OPF based on MapReduce approach.* Computer Communications, 2017. **98**: p. 52-71.
12. Wallgren, L., S. Raza, and T. Voigt, *Routing attacks and countermeasures in the RPL-based internet of things.* International Journal of Distributed Sensor Networks, 2013. **9**(8): p. 794326.
13. Chugh, K., L. Aboubaker, and J. Loo. *Case study of a black hole attack on LoWPAN-RPL*. in *Proc. of the Sixth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE), Rome, Italy (August 2012)*. 2012.
14. Kraus, R., et al., *Seven deadliest Microsoft attacks*. 2010: Elsevier.
15. Sonar, K. and H. Upadhyay, *A survey: DDOS attack on Internet of Things.* International Journal of Engineering Research and Development, 2014. **10**(11): p. 58-63.
16. Alhowaide, A., I. Alsmadi, and J. Tang, *Ensemble detection model for IoT IDS.* Internet of Things, 2021. **16**: p. 100435.
17. Shafiq, M., et al., *CorrAUC: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques.* IEEE Internet of Things Journal, 2020. **8**(5): p. 3242-3254.
18. Shen, Y., et al., *An ensemble method based on selection using bat algorithm for intrusion detection.* The Computer Journal, 2018. **61**(4): p. 526-538.
19. Midi, D., et al. *Kalis—A system for knowledge-driven adaptable intrusion detection for the Internet of Things*. in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 2017. IEEE.
20. Otoum, Y., D. Liu, and A. Nayak, *DL-IDS: a deep learning–based intrusion detection framework for securing IoT.* Transactions on Emerging Telecommunications Technologies, 2022. **33**(3): p. e3803.
21. Amouri, A., V.T. Alaparthy, and S.D. Morgera, *A machine learning based intrusion detection system for mobile Internet of Things.* Sensors, 2020. **20**(2): p. 461.
22. Amin, R., et al., *A light weight authentication protocol for IoT-enabled devices in distributed Cloud Computing environment.* Future Generation Computer Systems, 2018. **78**: p. 1005-1019.
23. Gope, P., et al., *Lightweight and privacy-preserving RFID authentication scheme for distributed IoT infrastructure with secure localization services for smart city environment.* Future Generation Computer Systems, 2018. **83**: p. 629-637.
24. Barcelo, M., et al., *IoT-cloud service optimization in next generation smart environments.* IEEE Journal on Selected Areas in Communications, 2016. **34**(12): p. 4077-4090.
25. Onah, J.O., et al., *Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment.* Machine Learning with Applications, 2021. **6**: p. 100156.
26. Moustafa, N., B. Turnbull, and K.-K.R. Choo, *An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things.* IEEE Internet of Things Journal, 2018. **6**(3): p. 4815-4830.
27. Li, Z., et al., *Nearfield target localization with a few snapshots for sonar array.* EURASIP Journal on Wireless Communications and Networking, 2018. **2018**(1): p. 1-7.
28. Kiran, K.S., et al., *Building a intrusion detection system for IoT environment using machine learning techniques.* Procedia Computer Science, 2020. **171**: p. 2372-2379.
29. Verma, A. and V. Ranga, *Machine learning based intrusion detection systems for IoT applications.* Wireless Personal Communications, 2020. **111**: p. 2287-2310.
30. Ge, M., et al. *Deep learning-based intrusion detection for IoT networks*. in *2019 IEEE 24th pacific rim international symposium on dependable computing (PRDC)*. 2019. IEEE.

31. Tian, Z., et al., *A distributed deep learning system for web attack detection on edge devices.* IEEE Transactions on Industrial Informatics, 2019. **16**(3): p. 1963-1971.
32. Khan, M.A., et al., *A deep learning-based intrusion detection system for mqtt enabled iot.* Sensors, 2021. **21**(21): p. 7016.
33. Ferrag, M.A., et al., *Deep learning-based intrusion detection for distributed denial of service attack in agriculture 4.0.* Electronics, 2021. **10**(11): p. 1257.
34. Lopez-Martin, M., et al., *Shallow neural network with kernel approximation for prediction problems in highly demanding data networks.* Expert Systems with Applications, 2019. **124**: p. 196-208.
35. Nimbalkar, P. and D. Kshirsagar, *Feature selection for intrusion detection system in Internet-of-Things (IoT).* ICT Express, 2021. **7**(2): p. 177-181.
36. Parimala, G. and R. Kayalvizhi. *An effective intrusion detection system for securing IoT using feature selection and deep learning*. in *2021 international conference on computer communication and informatics (ICCCI)*. 2021. IEEE.
37. Pang, J., et al., *Realizing the heterogeneity: A self-organized federated learning framework for IoT.* IEEE Internet of Things Journal, 2020. **8**(5): p. 3088-3098.
38. Cook, A.A., G. Mısırlı, and Z. Fan, *Anomaly detection for IoT time-series data: A survey.* IEEE Internet of Things Journal, 2019. **7**(7): p. 6481-6494.
39. Liang, W., et al., *Deep reinforcement learning for resource protection and real-time detection in IoT environment.* IEEE Internet of Things Journal, 2020. **7**(7): p. 6392-6401.
40. Tyagi, H. and R. Kumar, *Cloud computing for iot*, in *Internet of Things (IoT)*. 2020, Springer. p. 25-41.
41. Zorbas, D., et al., *Optimal Data Collection Time in LoRa Networks-A Time-Slotted Approach.* Sensors (Basel), 2021. **21**(4).
42. Shancang Li, L.D.X., and Xinheng Wang, *Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things.* IEEE Transactions on Industrial Informatics November 2013. **Volume: 9**(Issue: 4): p. Page(s): 2177 - 2186.
43. Luong, N.C., et al., *Data collection and wireless communication in Internet of Things (IoT) using economic analysis and pricing models: A survey.* IEEE Communications Surveys & Tutorials, 2016. **18**(4): p. 2546-2590.
44. Ali, Z.H., H.A. Ali, and M.M. Badawy, *Internet of Things (IoT): definitions, challenges and recent research directions.* International Journal of Computer Applications, 2015. **128**(1): p. 37-47.
45. Kiljander, J., et al., *Semantic interoperability architecture for pervasive computing and internet of things.* IEEE access, 2014. **2**: p. 856-873.
46. Noura, M., M. Atiquzzaman, and M. Gaedke, *Interoperability in internet of things: Taxonomies and open challenges.* Mobile networks and applications, 2019. **24**(3): p. 796-809.
47. Ganzha, M., et al., *Towards semantic interoperability between Internet of Things platforms*, in *Integration, interconnection, and interoperability of IoT systems*. 2018, Springer. p. 103-127.
48. Kim, J., J. Jeong, and J. Shin. *M2m: Imbalanced classification via major-to-minor translation*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
49. Shi, H., et al., *SAR Slow Moving Target Imaging Based on Over-Sampling Smooth Algorithm.* Chinese Journal of Electronics, 2017. **26**(4): p. 876-882.
50. Liu, Y., et al., *Missing value imputation for industrial IoT sensor data with large gaps.* IEEE Internet of Things Journal, 2020. **7**(8): p. 6855-6867.

51. Lin, W.-C. and C.-F. Tsai, *Missing value imputation: a review and analysis of the literature (2006–2017).* Artificial Intelligence Review, 2020. **53**(2): p. 1487-1509.
52. Khare, S. and M. Totaro. *Big data in IoT*. in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019. IEEE.
53. Chandrashekar, G. and F. Sahin, *A survey on feature selection methods*. Computers & Electrical Engineering, 2014. **40**(1): p. 16-28.
54. Mafarja, M., et al., *Augmented whale feature selection for IoT attacks: Structure, analysis and applications.* Future Generation Computer Systems, 2020. **112**: p. 18-40.

Article submitted 1 March 2023. Accepted at 30 April

Published at 30 Jun 2023.