Received: March 13, 2023. Revised: April 24, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Improved Intrusion Detection System Using Machine Learning with Singular Value Decomposition and Principal Component Analysis

Ruqaya A Al Hasan[1]* **Ekhlas Kadhum Hamza[1]**

[1]*Department of Computer Engineering, University of Technology, Iraq*
*Corresponding author's Email: Ekhlas.K.Hamza@uotechnology.edu.iq

Abstract: Among the most crucial components of a cyber physical system is a network of nodes via which a large number of autonomous, mobile or stationary sensors can communicate with one another. This network is known as a wireless sensor network (WSN). The network's nodes work together to sense the world around them, collect data on the items they detect, process that data, and then communicate it on to the network's owner since WSN has many potentials uses across many disciplines but little available resources, it is often coupled with IOT, making the network accessible to the outside world and susceptible to cyberattacks. Blackhole, grayhole, flooding, and scheduling attacks are some examples of common attacks in WSN that can do significant damage rapidly. Intrusion detection approaches for WSN suffer from issues like a low detection rate, a large calculation overhead, and a high false alarm rate because of the network's redundant and highly correlated data and the constraints imposed by sensor nodes' limited resources which is the research problem. This research proposes a solution, dubbed IDS-ML, that makes use of three different machine learning techniques—stochastic gradient descent (SGD), ridge regression (RR), and gaussian naive bayes—to solve the problem of intrusion detection in wireless sensor networks (GNB). In order to reduce the computational burden of the technique, principal component analysis (PCA) and singular value decomposition (SVD) are applied to the original traffic data to lower the feature space dimension. Once network threats have been identified, an IDS-ML model is utilized to categorize them. Based on the experiments with two datasets WSN-DS and UNSW-NB15, the proposed IDS-ML achieves significantly higher accuracy rate of 99% than state-of-the-art detection algorithms for WSN-DS and UNSW-NB15 dataset. As the achieved higher accuracy rate against normal, blackhole, grayhole, flooding, and TDMA attacks are 99%, 100%, 72%, 100%, and 78% respectively.

Keywords: Wireless sensor network, Intrusion detection, Machine learning, Accuracy, Internet of things.

1. Introduction

Wireless sensor networks (WSN) have advanced rapidly alongside the internet of things (IoT's) rise to prominence. This is because the internet of things has the ability to interconnect everything and progressively alter people's lifestyles[1, 2]. For the safety of IoT networks, stringent safeguards are essential. Security methods such as encryption, authentication, and other forms of protection have been introduced to ensure the WSN is secure. In contrast, the evolution of a wide variety of attack techniques has led to the emergence of threats that can evade traditional forms of security protection. If you're going to implement a WSN system on a massive scale, data security must be a top priority. More stringent security procedures are essential for protecting WSN systems [3, 4].

Unfortunately, WSN security cannot be guaranteed in full by passive Défense methods. A preventative protection technology must be made available [5]. One of the most useful active Défense technologies is the intrusion detection system (IDS) [6]. In the absence of conventional safeguards, data-driven IDS can proactively detect assaults.

However, as the volume of data transmitted over a network grows, it becomes more challenging for IDS to do analysis in real time. Also significant in weighing the pros and cons of IDS is the speed and efficiency with which data in WSN can be processed [7]. Abnormal network traffic exhibits feature such as

sudden traffic creation and unexpected parameter characteristics for the WSN. Normal network traffic is disrupted by attacks like wormholes, sinkholes, flooding, and jamming [8] because of the sheer volume and variety of data that makes up network traffic, the classifier can't quickly differentiate between typical and aberrant patterns [9]. Noise and other irrelevant characteristics in network traffic data make it hard to spot suspicious activity, necessitating more time and energy to investigate and reducing the likelihood of success [10].

In most cases, models and algorithms can only get rather close to the limits of machine learning, which are set by the data and features themselves. So, feature selection is crucial in machine learning. Datasets with tens of thousands of dimensions of feature space are becoming increasingly common as computing power is applied across more and more areas of human society. However, only a subset of characteristics can adequately capture the essence of things. The efficiency of machine learning algorithms is severely hindered because this subset of data is masked by an overwhelming quantity of irrelevant and duplicate features. So far, scientists have successfully integrated feature selection with machine learning, and the practice has seen extensive use in sectors like network traffic monitoring and security [11].

There needs to be a balance between a high accuracy and detection rate and a low amount of time spent on intrusion detection. Some methods of network intrusion detection are no longer applicable to WSN because of the many ways in which WSN differs from conventional computer networks. These differences include terminal kinds, data transmission, network topology, and many more. First, an IDS for WSN needs to be very accurate in identifying attackers, including unknown attacks. Second, an IDS for WSN needs to be fast and easy to use, ensuring little overhead on the WSN infrastructure [11]. An intrusion detection model specifically for WSNs is presented in this study. Key elements of this paper contribution are:

(1) The high volume and diversity of data that must be processed by the wireless sensor network are the root cause of the computationally costly intrusion detection approach and the poor detection performance of intrusion behaviour. Therefore, different feature selections, such as principal component analysis (PCA) and singular value decomposition (SVD), have been analysed as part of this study on intrusion detection in WSNs.
(2) In this study, three distinct classification strategies—Guessing naive bayes (GNB), ridge regression (RG), and stochastic gradient descent (SGD)—are evaluated and their classification abilities under WSN are compared to suggest an intrusion detection model named IDS-ML.
(3) The IDS-ML is used so that traffic attacks under WSN can be detected with fewer false positives. Traditional methods of intrusion detection in WSN have a number of drawbacks that this model attempts to remedy, including poor detection performance, slow real-time performance, and a high level of complexity. It's less resource intensive, more exact, and more robust.

The remaining sections of this work are structured as follows. Detection of Incursions in WSN is explained in section 2. Related research is discussed in section 3. the proposed intrusion detection technique for WSN is presented in section 4. As experimental settings are illustrated in section 5. Experiment outcomes and analyses are presented in section 6. The final section of this document outlines our plans for the future in section 7.

2. Detection of incursions in WSN

There are two main types of attacks that can be made against a WSN: passive and aggressive. The term "active attack" is used to describe destructive attacks. In In In this context, an opponent is anybody or anything that does harm to the system in a direct and immediate fashion. A passive attack, which does not interfere with the regular data connection, can be used to retrieve the effective data of the destination station sent by the source station. Damage to the network and compromised data security can result from an unauthorized access to valid information. There will be no disruption to the data transfer schedule due to a leak of sensitive information [12]. The attacker in a passive assault listens in on a conversation between two nodes, while in an active attack the attacker takes use of the broadcast nature of wireless communication media. The goal of an invasion can classify it as either an exterior or an inside invasion. Anyone can disrupt a WSN and steal sensitive data by using extremely powerful wireless receiving and transmission equipment. Common methods of these attacks include replay, injection, eavesdropping, and interference. Once a crucial node in the network, this inside invader has switched to offensive mode when its node was destroyed. Independent nodes, which use network resources independently of other nodes in the network and do not harm other nodes in the network directly [13], and malicious sensor nodes, which eavesdrop on, interfere with, or even control the communication of

the entire network by masquerading as normal nodes, are the two types of nodes that can launch an internal attack. Given the restrictions of the network's energy, processing power, communication bandwidth, and storage capacity, it is essential to tailor the architecture of the intrusion detection system to the particular requirements of the application scenario and environment design in WSNs.

3. Related work

With the proliferation of wireless LANs [14], notably Ad Hoc networks and wireless sensor networks, traditional wired network intrusion detection system (IDS) solutions are incompatible. This highlights the critical necessity for an intrusion detection system for wireless sensor networks. Intruder detection systems that use anomaly detection will look at any suspicious behavior. Researchers have used these findings to construct a variety of powerful anomaly detection systems, most of which are variants on artificial immunity algorithms, clustering algorithms, machine learning algorithms, and statistical learning models.

To deal with these problems, Yao et.al, [15] provide a multilevel framework for intrusion detection models called multilevel semi-supervised ML (MSML). First, there is pure cluster extraction; second, there is pattern discovery; third, there is fine-grained classification (FC); and fourth, there is model update. We define a "pure cluster" and present a hierarchical semi-supervised k-means approach to locate all pure clusters in the pure cluster section.In order to effectively detect LDoS attacks in WSN, Chen et al. [16] devised a Hilbert Huang Transformation (HHT) technique that uses joint analysis. The primary objective of this work was to build a reliable attack detection framework by making use of the nodes with higher trust value. Its primary goal was to achieve lower energy use, travel time, and traffic volume. To better protect WSNs, Hu et al. [17] combined the cuckoo search optimization (CSO) method with the support vector machine (SVM) classification method. Predicting intrusions from provided network datasets with high classification efficiency and performance rate was the primary focus of this effort. To achieve this goal, we used the map reduction technique to efficiently parallelize the SVM classification model's parameters over several nodes. However, this study is constrained by issues such as a low accuracy, a slow response time, and a high misclassification rate.

To identify anomalies in the NSL-KDD dataset, Liu et al. [18] used an EM approach to expectation maximization. In this paper, we looked at several distinct kinds of attacks, including synflood, land, ping of death, sweeping, and UDP flood. To achieve smart, sustainable energy management, Hemanand et al. [19] suggested applying the existing glow worm swarm optimization technique across IoT sensors to detect the devices in need of energy and distribute appropriate energy on a need basis. According to Jayalakshmi et al. [20], the routing protocol should be one of the factors examined when gauging a network's efficacy. It was proposed by Gopalakrishnan et al. that the security of the system may be enhanced by deploying highly secured cryptographic algorithms on each node in the network. To mitigate the wide-ranging effects of denial-of-service (DoS) attacks while keeping energy consumption to a minimum, feature selection models for NIDSs are proposed by Almomani [8]. Particle swarm optimization (PSO), grey wolf optimization (GWO), firefly algorithm (FFA), and the genetic algorithm (GA) all form the basis of this concept (GA). The proposed model is made with the intention of enhancing NIDS functionality. The suggested model uses Anaconda Python open source's wrapper-based methods with the GA, PSO, GWO, and FFA algorithms to pick features, as well as filtering-based methods for the mutual information (MI) of the GA, PSO, GWO, and FFA algorithms, which yielded 13 sets of rules. Support vector machine (SVM) and J48 ML classifiers are used on the UNSW-NB15 dataset to assess the proposed model's output characteristics. Feature selection models for NIDSs are proposed by Almomani [8]. Particle swarm optimization (PSO), grey wolf optimization (GWO), firefly algorithm (FFA), and the genetic algorithm (GA) all form the basis of this concept. The proposed model is made with the intention of enhancing NIDS functionality. The suggested model uses Anaconda Python open source's wrapper-based methods with the GA, PSO, GWO, and FFA algorithms to pick features, as well as filtering-based methods for the mutual information (MI) of the GA, PSO, GWO, and FFA algorithms, which yielded 13 sets of rules. Support vector machine (SVM) and J48 ML classifiers are used on the UNSW-NB15 dataset to assess the proposed model's output characteristics.

MQTTset, presented by Vaccari et al. [21], is a dataset dedicated to the MQTT protocol, which is commonly used in IoT networks. By combining the official dataset with cyberattacks against the MQTT network, we show the creation of the dataset and validate it through the definition of a hypothetical detection system. The obtained results show how machine learning models may be trained using MQTTset to create detection systems that can secure IoT environments.

Table 1. Review on existing strategies

Research	Year	Dataset	Algorithm	Attacks executed	Accuracy	Precision	Recall	F1 Score
Yao et.al, [15]	2019	KDDCUP99 dataset	a hierarchical semi-supervised k-means algorithm (HSK-means)	Normal DOS Probe U2R R2L	N/A	96 100 89 76 65	86 100 98 14 4	92 100 74 76 94
Hu et al. [17]	2019	NSL-KDD dataset	CS - SVM MR-SVM	N/A	95.97 96.16	90.48 88.41	61.29 65.69	73.08 74.31
Chen et al. [16]	2019	N/A	Hilbert–Huang transform (HHT)	Random routing REQuest (RREQ) Low-rate Denial of Service (LDoS)	N/A	N/A	N/A	N/A
Liu et al. [18]	2020	NSL-KDD dataset	RF DT Bagging SVM NB BN AdaBoost XGBoost	Syn Flood Land UDP Flood Ping of Death (PoD) Smurf IP Sweeping Port Scan	0.966 0.996 0.967 0.957 0.452 0.882 0.740 0.970	0.969 0.969 0.969 0.948 0.904 0.944 0.663 0.970	0.967 0.967 0.967 0.957 0.452 0.882 0.740 0.968	0.968 0.968 0.968 0.951 0.545 0.902 0.646 0.968
Vaccari et.al,[21]	2020	MQTTset Message (Queue Telemetry) Transport	Neural network, random forest, Naïve Bayes, Decision tree, Gradient boost, Multilayer perceptron	flooding denial of service, MQTT Publish Flood, SlowITe malformed Data, brute force authentication	0.993268 0.994299 0.987903 0.977972 0.991131 0.94688	N/A	N/A	0.9932 0.9943 0.9897 0.9850 0.9916 0.9636
Kumar et.al, [22]	2020	UNSW-NB15 and real time data set at NIT Patna CSE lab (RTNITP18)	Different decision tree models (C5, CHAID, CART, QUEST) are trained with selected 13 features of the dataset	Exploit, DOS, Probe, Generic and Normal	high	69.9 50.37 99.21 99.7 81.17	54.6 5 71.7 96.7 98	high
Almomani et. al, [25]	2020	UNSW-NB15 dataset.	the support vector machine (SVM) and J48 ML classifiers	----------	90.119 90.484	N/A	N/A	N/A
Makhija et.al,[23]	2022	MQTTset (Message Queue Telemetry Transport)	RF, KNN ,and SVM classifier	unauthorized access, denial of service, packet sniffing, and malware injection	96	N/A	N/A	N/A
Hemanand et. al[24]	2022	NSL-KDD and UNSW-NB15	Cuckoo Search Greedy Optimization (CSGO) and Likelihood Support Vector Machine (LSVM)	Probe, DOS, U2R, R2L	99.65	99.99	98.69	99.5

A innovative misuse-based intrusion detection system is proposed by Kumar et.al, [22], which can identify five types of attacks in a network: exploit, DOS, probe, generic, and normal. In addition, the KDD99 or NSL-KDD 99 data set is used in the majority of the works that are similar to IDS. When it comes to detecting modern threats, these data sets are now regarded useless and antiquated. In this paper, we use the UNSW-NB15 dataset as an offline resource for developing our own integrated classification-based algorithm for sniffing out cybercrime.

The effectiveness of an attack on a MQTT-based IoT system may be predicted using a number of different machine learning models, as demonstrated by the work of Makhija et.al, [23]. To compare the efficacy of the models, we employed the precision, accuracy, and F1 score as evaluation criteria. Results demonstrated that random forest's performance was very accurate, with a 96 percent degree of certainty.

Using the Cuckoo search greedy optimization (CSGO) and likelihood support vector machine (LSVM) models, Hemanand et. Al [24] proposed work creates an intelligent IDS system for improving WSN security. This model takes into account the most popular network datasets for validation, including NSL-KDD and UNSW-NB15. At first, the attributes are normalized via dataset preprocessing via the elimination of extraneous data, the prediction of missing values, and the application of filters. In order to pick the optimum features, the CSGO algorithm must be fed the optimal number of features that were determined during preprocessing. The final step is to forecast the categorized label as normal or abnormal using a machine learning classification technique based on the linear support vector machine (LSVM). During the results evaluation process, many performance measurements are used to verify and compare the effectiveness of the suggested security model.

4. WSN intrusion detection architecture

The data-gathering module, the detection model, and the reaction module are the three main pillars of WSN intrusion detection technology. The environment is scoured for information by the information collection module, which then passes the information along to the detection subsystem. To assess if an intrusion has occurred within the WSN, an analyzer is housed within the detection module and performs a thorough examination and analysis of the collected data traffic information. If something out of the ordinary is discovered, the detection module will notify the response module. As shown in Fig. 1, WSN can be used for intrusion detection. Network nodes can be represented by the sensor node (SN), the cluster head (CH), or the sink node (Sink) [14, 26].

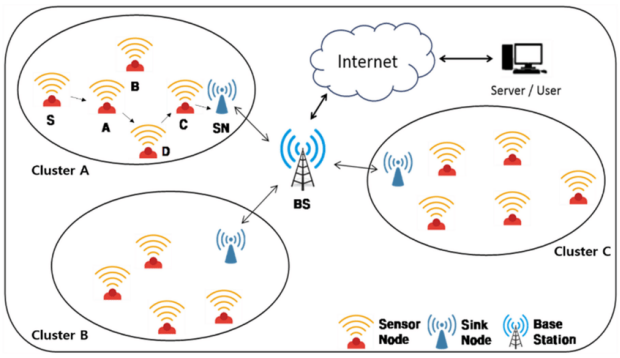

Figure. 1 The process for detecting intrusions in wireless sensor networks [27]

Distributed intrusion detection is used in this network to lessen the communication load and save on energy. In distributed detection, the cluster leader is responsible for worldwide organization and coordination of computer tasks. A portion of the computing cost of the cluster head can be offloaded to regular sensor nodes, and communication overhead can be reduced by relaying less information to the cluster head [28]. In order to obtain high detection and accuracy for IDS under WSN, many researchers turn to increasingly complex data mining methods. However, the substantial processing overhead of such methods renders their real-time deployment in wireless sensor networks impracticable. The high computational cost of IDS can be traced back to several factors, the most prominent of which are the huge feature dimensions of the input data, an excess of redundant data, and a lack of sufficient data preparation.

Feature selection is a method for narrowing down a large set of candidate features to a more manageable number. Therefore, in terms of data pre-processing, considering the importance of the data itself, principal component analysis (PCA) and singular value decomposition (SVD) are employed to maximize the preservation of information, boost the accuracy of the classification algorithm's detection, and lessen the computational burden of IDS. Using this technique, the data analysis module shown in Fig. 1 can receive event information and evaluate it to determine if the behaviour being seen is an incursion. An overview of the algorithmic framework is shown in Figs. 2 and 3.

5. The proposed research methodology

The architectural layout of the system that is suggested by this research may be found illustrated in

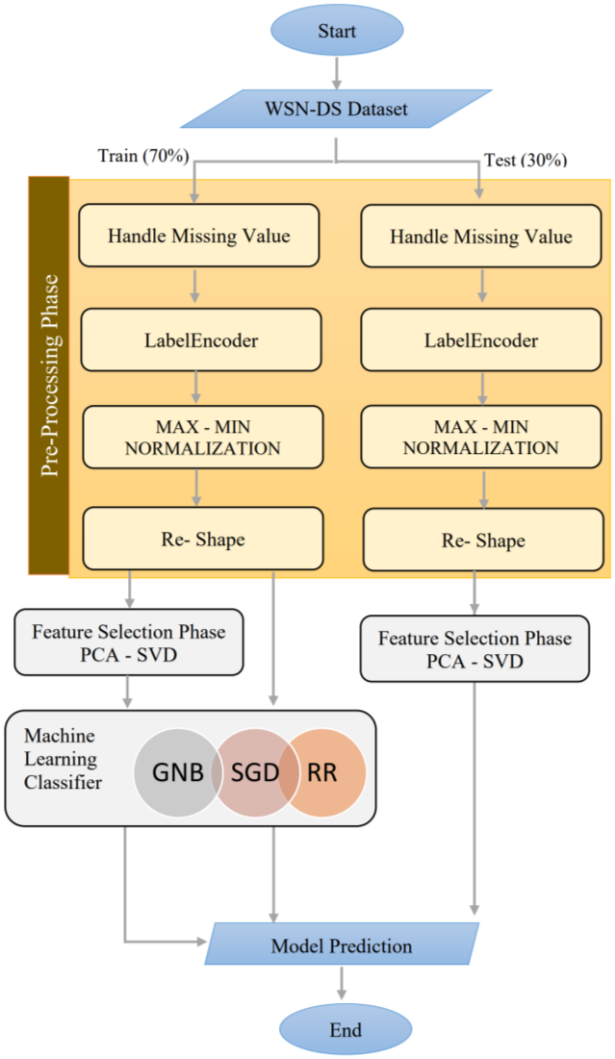

Figure. 2 Specification of the algorithm framework for WSN-DS dataset

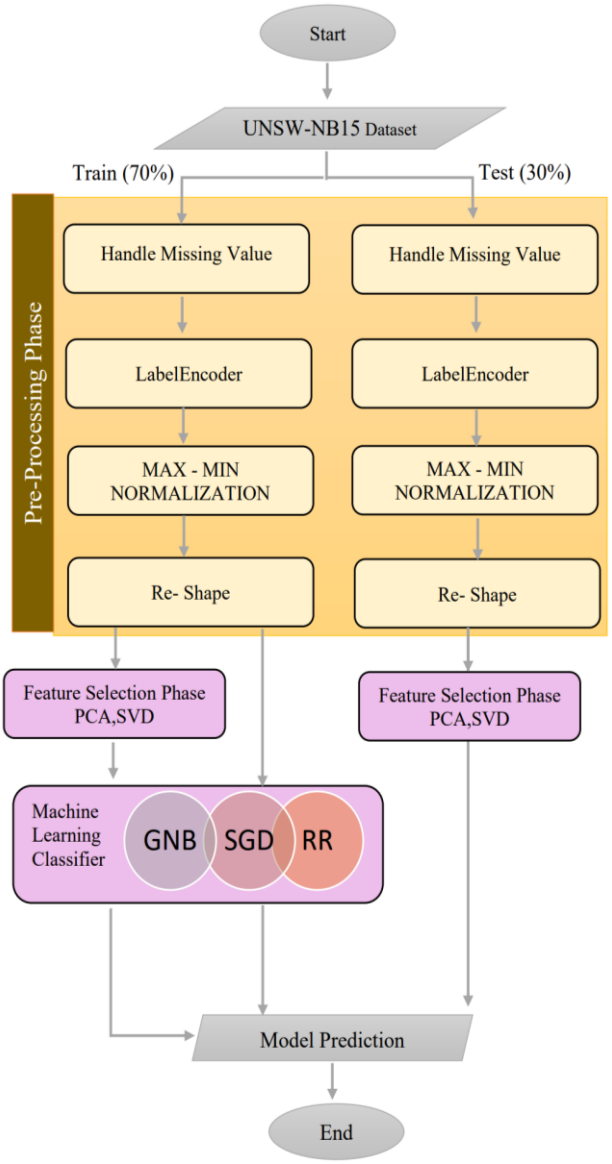

Figure. 3 Specification of the algorithm framework for UNSW-NB15 dataset

Figs. 2 and 3. The first section is dedicated to the processing of data in its many forms.

This method is frequently referred to as the data engineering process. This is a necessary stage for any kind of learning to take place. Cleaning, normalization, and feature selection are the three stages of data processing. To choose the most relevant features, a filter-based approach that takes cues from principal component analysis and singular value decomposition is employed. Model training using the training set comes after the selection of the necessary feature vector. Once a model has been trained, it can be checked against the validation set. At last, the model that has been validated is put to the test on the test dataset.

Table 1. attack-type-characteristic-value conversion table

Original eigenvalue	Transformed eigenvalue
Normal	0
Grayhole	1
Blackhole	2
TDMA	3
Flooding	4

5.1 The preparation of data

1) Collecting and mapping information

The sample data's label feature is a string of letters; to remove it from the algorithm, and must translate those characters into numbers. Normal, blackhole, grayhole, flooding, and TDMA are the five data types that make up the Attack category. Due to the incalculability of such information, the ordinal digits 0, 1, 2, 3, and 4 to arrange the data un a sequential order are employed. Alteration in

accordance with Table 1.

2) LabelEncoder

The string representation of labels for nominal and ordinal features in a collection of categories. It's possible that some labels will contain ordinal characteristics (ordering) while others won't (nominal features). In order to ensure that the learning algorithm correctly understands the features, it is crucial that labels be encoded in numerical form during the data pre-processing phase. Labels are given numeric values in LabelEncoder's encoding process.

3) Maximum and minimum normalization

Since the range of characteristics in the data, from less than 1 to hundreds of thousands, has a significant effect on several classification algorithms, and must normalize the continuous data. The extreme values from Eq. (1) are used here for normalization. Where xj is the raw data for the j-dimensional feature, Minj represents the absolute minimum value for the feature, Maxj represents the absolute highest value for the feature, and x j represents the normalized data for the feature.

$$x_j^* = \frac{x_j - Min_j}{Max_j - Min_j} \quad (1)$$

5.2 Features extraction

1) Principal component analysis

Discovering patterns in high dimensional data is a prominent use case for principal component analysis (PCA). Using a smaller set of typical feature photos (called Eigenobject) to represent both known and unfamiliar faces is the information theory method driving PCA's objective. The statistical evidence for principal component analysis's (PCA) usage in face recognition technology demonstrates its utility for identifying and validating facial traits. To use the PCA method, a matrix of facial images in two dimensions must be transformed into a vector in one dimension. It makes no difference whether a one-dimensional vector is oriented along a row or a column [22, 23].

2) Singular value decomposition

Similarly, singular value decomposition (SVD) can be used to partition a dataset. It has several applications in signal processing and statistics, including feature extraction, matrix approximation, and pattern identification. PCA, however, is unable to extract features from a single signal, nor can it reveal information about the features present in a signal with changing frequencies. Because frequency differences might obscure true differences between physiological states, SVD can be a more effective tool for feature extraction than principal component analysis [29-31].

5.3 Classification model

For intrusion detection, the IDS-ML classification method on data collected via wireless sensor networks and filtered using the sequence backward feature selection approach is employed. Gradient based techniques constitute the basis of IDS-ML, a fast, distributed, high-performance gradient boosting framework [32]. The fundamental component of IDS-ML is a variant of the histogram technique that reduces the number of features and the number of samples used in each training session.

1. Gaussian naive bayes

Using a probabilistic strategy and the Gaussian distribution, the machine learning (ML) classification method known as Gaussian naive bayes (GNB) can be applied. Each parameter (also known as a feature or predictor) in Gaussian naive bayes is treated as though it were capable of predicting the output variable on its own. The combined forecast for all factors yields a final prediction, which in turn yields a probability for the dependent variable to be classified in each group. If two groups have equal probabilities, the one with the higher probability wins. A Gaussian distribution is assumed for the feature probabilities as in Eq. (2):

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

For each Y class, the formulas above calculate the variance and mean of the continuous variable X using the symbols σ and μ, respectively. A Gaussian naive bayes (GNB) classifier is depicted in action in Fig. 4. Each piece of information is then placed in the category to which it is most closely related. However, the GNB considers not just the distance from the mean but also how this compares to the class variance when computing this proximity [32].

2. Stochastic gradient descent

Specifically for linear classifiers, the model is a powerful learning algorithm. A less precise estimate of the gradient can stand in for the real thing. The stochastic (or "operational") gradient descent algorithm estimates the gradient of the cost function by assigning a gradient to each learning element. Several parameter shifts were made to account for the projected variations. When a new piece of data was introduced for learning, the model's parameters were recalculated. The stochastic gradient descent method

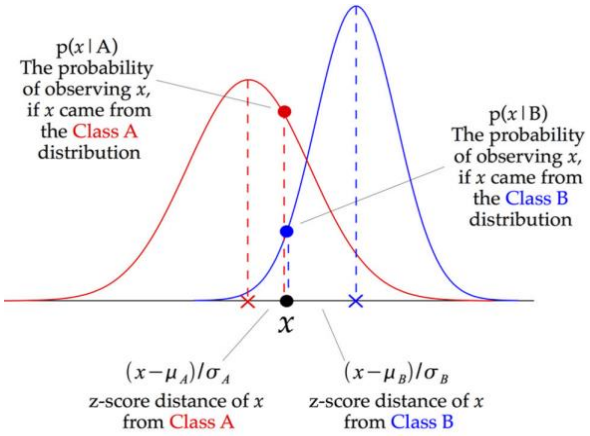

Figure. 4 A graphical representation of the operation of a naive bayes classifier

significantly outperforms the traditional method on large datasets [33]. This model is an effective form of facilitation. Easily digestible SGD updates as follows in Eq. (3) :

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_t \nabla l_i\left(\theta^{(t)}\right) \qquad (3)$$

In this equation, t stands for the number of iterations, and both and reflect the size of the learning set used to adjust the parameters. Here, the index I will have a new value chosen at random before each repetition. In actual reality, it is common to randomly mix up the samples before analyzing them [29].

3. Ridge regression

In order to analyze data that is affected by multicollinearity, ridge regression is employed as a model tuning technique. The L2 regularization is executed by this technique. Predicted values are off by a significant margin when multicollinearity occurs because least-squares are unbiased and variances are high. The ridge regression model is a variant on the standard regression equation that includes a correction function, as shown in Eq (4).

$$SSE_{L2} = \sum_{i=1}^{n} \left(yi - y_i^{\wedge}\right)^2 + \lambda \sum_j^{\rho} \beta_j^2 \qquad (4)$$

For a traditional regression, the left side of the equation is the key. On the right, the square root of each beta number is seen. Also, these numbers sum up to something. Multiplying by the adjustment parameter then produces a standardization factor for the model. The value of is crucial in ridge regression. It enables regulating the relative importance of the two terms. In this case, then, the punishment sentence is really. The ridge regression function has an alpha parameter that stands for. The penalty can be adjusted by modifying the alpha value. The standard regression equation is obtained if =0. Therefore, the cost is increased as Alpha rises. In this way, the coefficients become smaller [34].

Table 2. Class label description of WSN-DS dataset

Class	Description
Normal	Normal connection records
Blackhole	It is a kind of 'DoS' attack where an attacker attacksLEACH protocol and during initial time itself they publicize themselves as a CH
Grayhole	It is a kind of 'DoS' attack where an attacker attacksLEACH protocol and during initial time itself they publicize themselves as a CH for other nodes
Flooding	Using different ways, an attacker attacks LEACH protocol
Scheduling	Scheduling attack happens during the setup phase of LEACH protocol

Table 3. Feature description of UNSW-NB15 dataset

No.	Feature	Category	No.	Feature	Category
f1	dur	float	f22	dtcpb	integer
f2	proto	nominal	f23	dwin	integer
f3	service	nominal	f24	tcprtt	float
f4	state	nominal	f25	synack	float
f5	spkts	integer	f26	ackdat	float
f6	dpkts	integer	f27	smean	integer
f7	sbytes	integer	f28	dmean	integer
f8	dbytes	integer	f29	trans_depth	integer
f9	rate	float	f30	response_body_len	integer
f10	sttl	integer	f31	ct_srv_src	integer
f11	dttl	integer	f32	ct_state_ttl	integer
f12	sload	float	f33	ct_dst_ltm	integer
f13	dload	float	f34	ct_src_dport_ltm	integer
f14	sloss	integer	f35	ct_dst_sport_ltm	integer
f15	dloss	integer	f36	ct_dst_src_ltm	integer
f16	sinpkt	float	f37	is_ftp_login	binary
f17	dinpkt	float	f38	ct_ftp_cmd	integer
f18	sjit	float	f39	ct_flw_http_mthd	integer
f19	djit	float	f40	ct_src_ltm	integer
f20	swin	integer	f41	ct_srv_dst	integer
f21	stcpb	integer	f42	is_sm_ips_ports	binary

6. Experimental settings

Here, the publicly available WSN-DS dataset was employed [35] for the experiment. created specifically for use with WSNs, this dataset contains information used to detect intrusions (WSN). Blackhole, grayhole, flooding, and scheduling are the four forms of DoS assaults seen in WSN-DS. Table 2

contains the comprehensive statistical data. Of the total number of samples in both the training and testing sets, 224796 (or 70%) were drawn at random from the former, and 149865 (or 30%) from the latter.

The experiments were also conducted with data from the UNSW-NB15 assaults dataset [36]. As may be shown in Table 3, the 42 elements present in the UNSW-uncluttered NB15's design. There are a total of 42 features, of which only three are not numerical in nature (categorical features).

But the confusion matrix (CM) is utilized to evaluate the accuracy, recall, precision, and F-measure of our approach on the dataset. In Eqs. (5) - (8), the true positive (TP) and false negative (TN) counts are balanced by the false positive (FP) and false negative (FN) counts, respectively[37-39]. precision: the percentage of instances that are accurately classified; Take into account again the percentage of "good" components that were properly assigned to the "good" group; Accuracy: the percentage of false alarms that occur when using a detection model that initially misclassified some components as false positives; The F-score is the mean.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F_1 = 2 * \frac{precision*recall}{precision+recall} = \frac{2TP}{2TP+FP+FN} \quad (8)$$

7. Result analysis and discussion

The effectiveness of the suggested model is assessed here. The researchers performed several distinct experiments: 1) Examining IDS-ML with and without feature extraction phase (PCA/SVD with 10 or 15 features) in contrast to machine learning classification methods. 2) Examining IDS-ML in comparison to other machine learning classification techniques; 3) Examining IDS-ML in light of four different assaults accuracy and recall rates for detection to be high enough for network use and eventual integration with the network's intrusion detection system. Traditional classification techniques, including Gaussian naive bayes (GNB), ridge classifier (RG), and stochastic gradient descent (SGD) on WSN-DS dataset without feature selection phase, are compared in Table 4.

While Table 5 illustrates results on UNSW-NB15 dataset without feature selection phase. When compared to other algorithms, the IDS-ML classification algorithm excels in terms of accuracy, precision, recall, and the F-measure.

Table 4. Comparison of multiple classification measurements on WSN-DS dataset without feature selection phase

Algorithms			Measures		
		Accuracy	Precision	Recall	F1-score
IDS-ML	RG	0.95	0.96	0.96	0.96
	GNB	0.95	0.95	0.95	0.95
	SGD	0.96	0.97	0.96	0.97

Table 5. Comparison of multiple classification measurements on UNSW-NB15 dataset without feature selection phase

Algorithms			Measures		
		Accuracy	Precision	Recall	F1-score
IDS-ML	RG	0.89	0.98	0.89	0.92
	GNB	0.90	0.91	0.9	0.9
	SGD	0.87	0.99	0.87	0.92

For better estimator accuracy or better performance on very high-dimensional datasets, the feature selection module can be used to perform feature selection/dimensionality reduction on sample sets. This paper uses two algorithm PCA and SVD with 10 or 15 features. Figs. 5-8 the results obtained from applying feature selection with ML algorithms on WSN-DS dataset.

Figs. 9-12 shows the findings that were achieved by applying feature selection using ML algorithms to the UNSW-NB15 dataset via PCA and SVD with either 10 or 15 features.

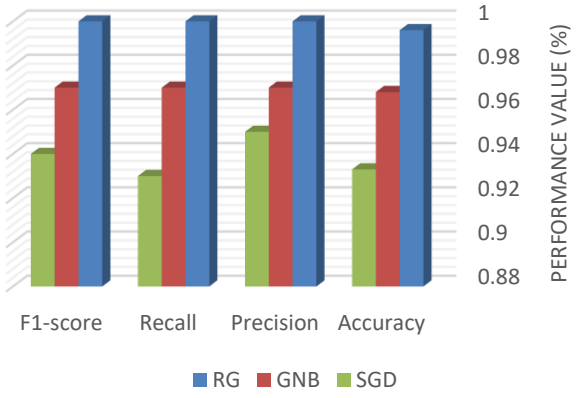

Figure. 5 Comparison of multiple classification measurements with (PCA10) feature selection on WSN-DS

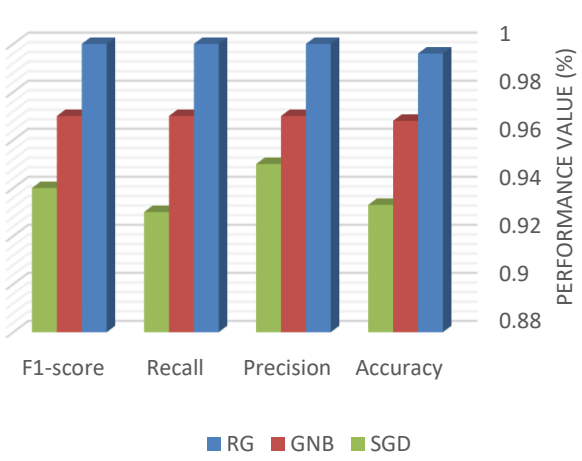

Figure. 6 Comparison of multiple classification measurements with (PCA15) feature selection on WSN-DS

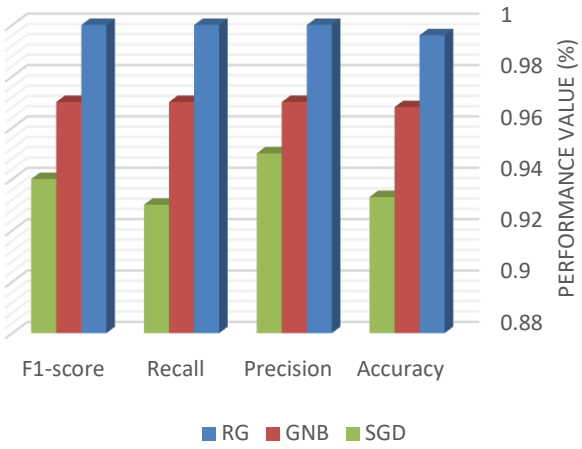

Figure. 9 Comparison of multiple classification measurements with (PCA 10) feature selection on UNSW-NB15 dataset

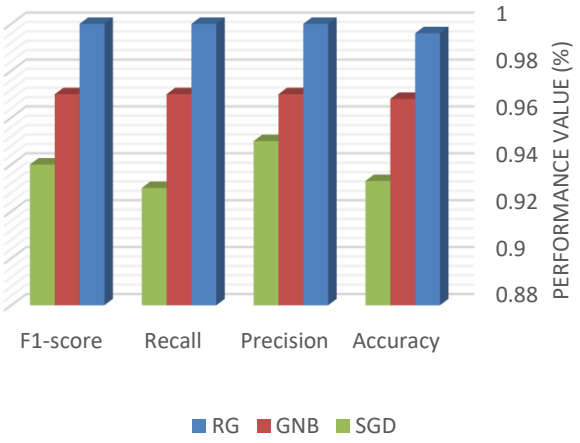

Figure. 7 Comparison of multiple classification measurements with (SVD 10) feature selection on WSN-DS

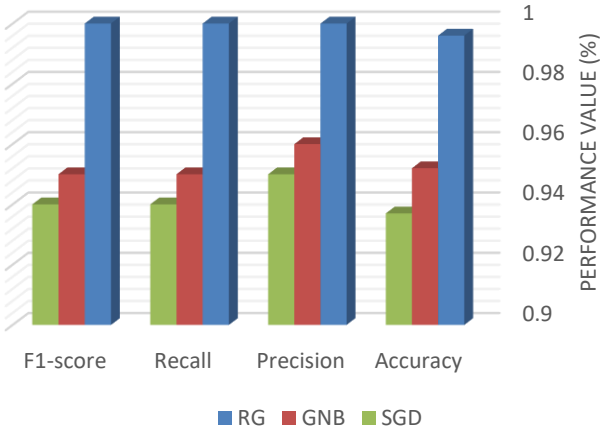

Figure. 10 Comparison of multiple classification measurements with (PCA 15) feature selection on UNSW-NB15 dataset

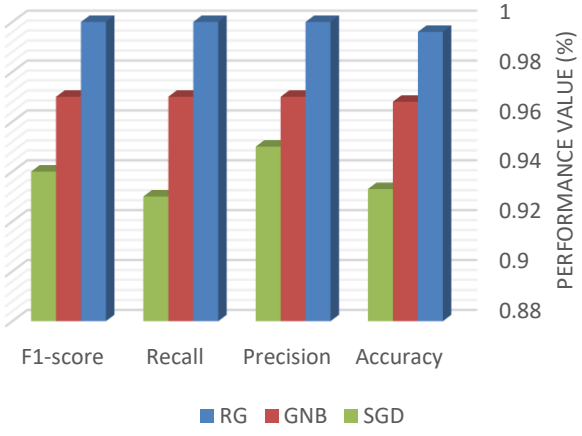

Figure. 8 Comparison of multiple classification measurements with (SVD 15) feature selection on WSN-DS

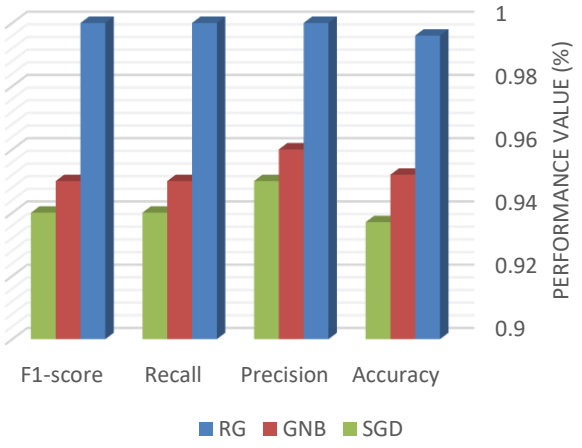

Figure. 11 Comparison of multiple classification measurements with (SVD 10) feature selection on UNSW-NB15 dataset

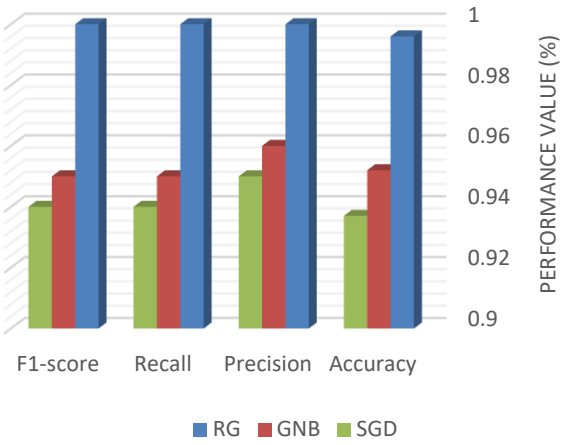

Figure. 12 Comparison of multiple classification measurements with (SVD 15) feature selection on UNSW-NB15 dataset

Table 6. Comparison of IDS-ML algorithm and other methods on WSN-DS

Algorithms		Measures				
		Normal	Grayhole	Blackhole	TMDA	Flooding
ANN [40]		0.998	0.756	0.928	0.922	0.994
DNN [41]		0.98	0.919	0.939	0.992	0.994
J48 [35]		0.999	0.982	0.993	0.927	0.975
SMO [35]		0.994	0.955	0.955	0.862	0.941
IDS-ML	**RG**	**0.65**	**0.72**	**0.47**	**0.65**	**0.71**
	GNB	**0.97**	**0.61**	**1.00**	**0.63**	**1.00**
	SGD	**0.99**	**0.42**	**0.94**	**0.78**	**0.83**

Table 7. Comparison of IDS-ML algorithm and other methods on UNSW-NB15 dataset

Algorithms	Feature Extraction Technique	Accuracy
CS - SVM [17]	SVM	95.97
MR-SVM [17]	SVM	96.16
RF [21]	N/A	99.42
J48 [25]	N/A	90.484
RF [23]	SVM	96
LSVM [24]	CSGO	99.65
CNN-BiLSTM [42]	O-SS-SMOS	77.16
IDS-ML	PCA10	99.06
IDS-ML	PCA15	99.66
IDS-ML	SVD10	99.16
IDS-ML	SVD15	99.26

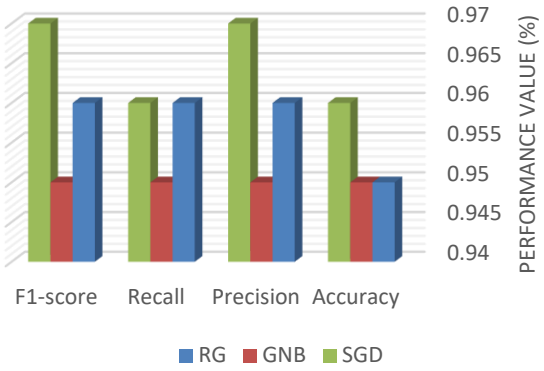

Figure. 13 Evaluation of several classification metrics

To compare the suggested algorithms to the state-of-the-art, ran them through the Gaussian nave bayes (GNB), ridge classifier (RG), and stochastic gradient descent (SGD) feature selection methods; the feature ranking for these algorithms is shown in Table 6. The proposed model has greater performance. The outcomes of the experiments are detailed in Table 6, as well as in Fig. 13. In order to evaluate the classification performance of our system, and make use of the confusion matrix, abbreviated as CM.

Table 7 contrasts the strategies suggested in this paper with those found in the literature review. According to the findings, when comparing ML methods for the feature selection scheme, IDS-ML performed better than any of the others, and when comparing ML methods for the multiclass configuration, it was the best option.

After conducting these studies, the researchers have determined the following:

The results shown in Tables 4 and 5 demonstrate that the SGD method achieves superior accuracy and recall compared to the other algorithms tested on the wireless sensor network dataset WSN-DS. The sample weight can be reflected in the size of the gradient if the SGD algorithm is being used. In general, a model's accuracy may be judged by its gradient, and a smaller gradient suggests greater accuracy.

As can be shown in Figs. 5-12, the accuracy, F-measure, and other indicators of the algorithm are impacted by the decreasing of the feature dimension that occurs following feature selection of the data. The feature selection algorithm stands out as the most effective of the three approaches. When dealing with WSN-DS and UNSW-NB15 data, it is essential to have the capability to take feature dependencies into consideration as well as the interplay between feature subset search and model selection. Because the other three methods do not take into account the classifier's interaction with the data, it is simple to eliminate

certain redundant internally dependent characteristics. However, the discrimination performance of certain features is poor when the data containing those features is processed as a whole, yet the features themselves offer tremendous potential for discrimination. The learning algorithm of the wrapper, which makes use of prediction accuracy, is responsible for calculating the benefits and downsides of the subset that has been selected. The ability to select a subset of characteristics that will be helpful during the process of learning by combining the use of classifiers with the practice of feature selection.

When compared to other approaches like MR-SVM [17], RF [21], J48 [25], LSVM [24] , CNN-BiLSTM [42] and RF [23], IDS-detection ML's performance is shown to be superior in Table 6. This is because initially it chooses features from the traffic data collected by the sensor nodes. In order to lower the feature dimension of the traffic. To accomplish this goal of lowering the dimensionality of the traffic data while simultaneously increasing the model's precision, an approach based on PCA and SVD has been implemented. Low detection performance, poor real-time detection, and excessive model complexity are only some of the issues plaguing current approaches to feature selection and classification in wireless sensor network intrusion detection systems. All of these issues can be resolved with this approach because they are dealt with separately. The model's high real-time performance and robust detection skills aid in keeping it from getting over-fit.

8. Conclusion

Currently, the most common method for detecting malicious software is a combination of feature selection algorithms and machine learning techniques. The model improves in generalization and overfitting is minimized when the number of features and the dimensionality are both decreased via the feature selection procedure. In contrast, it can help clarify the relationship between features and their associated values. First, run trials using several different machine learning techniques. When compared to its competitors, IDS-ML is a huge step up in terms of accuracy. Because it is a decision-based learning algorithm within a gradient boosting framework, which allows for faster training efficiency, lower memory use, a greater accuracy of 99%, and the ability to process massive amounts of data, it is particularly well-suited to these tasks. Different algorithms for intrusion detection in WSNs are compared with one another in an effort to boost IDS-performance ML even further. Maximal feature extraction during data pre-processing reduces dimensionality and eliminates data redundancy, eliminating IDS's high processing cost. It eliminates the issue. Next, increase precision and recall with IDS-ML. This system has a high detection rate, low false alarms, and little calculation, according to experiments and studies on similar mechanisms. In wireless sensor networks, it detects intrusions.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

The first author wrote the methodology, software, and formal analysis, while the second author reviewed, edited, and supervised.

References

[1] T. H. Nasser, E. K. Hamza, and A. M. Hasan, “MOCAB/HEFT algorithm of multi radio wireless communication improved achievement assessment”, *Bulletin of Electrical Engineering and Informatics*, Vol. 12, No. 1, pp. 224–231, 2023, doi: 10.11591/eei.v12i1.4078.

[2] M. Zhou, Y. Wang, Z. Tian, Y. Lian, Y. Wang, and B. Wang, “Calibrated Data Simplification for Energy-Efficient Location Sensing in Internet of Things”, *IEEE Internet Things J.*, Vol. 6, No. 4, 2019, doi: 10.1109/JIOT.2018.2869671.

[3] M. A. Hussein and E. K. Hamza, “Secure Mechanism Applied to Big Data for IIoT by Using Security Event and Information Management System (SIEM)”, *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 6, pp. 667–681, 2022, doi: 10.22266/ijies2022.1231.59.

[4] E. K. Hamza and S. N. Jaafar, “Nanotechnology Application for Wireless Communication System”, in *Materials Horizons: From Nature to Nanomaterials*, 2022, doi: 10.1007/978-981-16-6022-1_6.

[5] I. Batra, S. Verma, Kavita, and M. Alazab, “A lightweight IoT-based security framework for inventory automation using wireless sensor network”, *International Journal of Communication Systems*, Vol. 33, No. 4, 2020, doi: 10.1002/dac.4228.

[6] O. A. Osanaiye, A. S. Alfa, and G. P. Hancke, “Denial of Service Defence for Resource Availability in Wireless Sensor Networks”, *IEEE Access*, Vol. 6, 2018, doi: 10.1109/ACCESS.2018.2793841.

[7] A. Y. Hussein, A. T. Sadiq, and A. T., "Meerkat Clan-Based Feature Selection in Random Forest Algorithm for IoT Intrusion Detection", *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, Vol. 22, No. 3, 2022, doi: 10.33103/uot.ijccce.22.3.2.

[8] T. T. H. Le, T. Park, D. Cho, and H. Kim, "An Effective Classification for DoS Attacks in Wireless Sensor Networks", in *International Conference on Ubiquitous and Future Networks, ICUFN*, 2018, doi: 10.1109/ICUFN.2018.8436999.

[9] D. Selvamani and V. Selvi, "A Comparative Study on the Feature Selection Techniques for Intrusion Detection System", *Asian Journal of Computer Science and Technology*, Vol. 8, No. 1, 2019, doi: 10.51983/ajcst-2019.8.1.2120.

[10] P. Li, W. Zhao, Q. Liu, X. Liu, and L. Yu, "Poisoning machine learning based wireless IDSs via stealing learning model", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, doi: 10.1007/978-3-319-94268-1_22.

[11] S. K. Pandey, "An anomaly detection technique-based intrusion detection system for wireless sensor network", *International Journal of Wireless and Mobile Computing*, Vol. 17, No. 4, 2019, doi: 10.1504/IJWMC.2019.103110.

[12] G. Liu, H. Zhao, F. Fan, G. Liu, Q. Xu, and S. Nazir, "An Enhanced Intrusion Detection Model Based on Improved kNN in WSNs", *Sensors*, Vol. 22, No. 4, 2022, doi: 10.3390/s22041407.

[13] P. Michiardi and R. Molva, "Core: A Collaborative Reputation Mechanism to Enforce Node Cooperation in Mobile Ad Hoc Networks", *IFIP Advances in Information and Communication Technology*, pp. 107–121, 2002, doi: 10.1007/978-0-387-35612-9_9.

[14] M. Zhou, Y. Liu, Y. Wang, and Z. Tian, "Anonymous crowdsourcing-based WLAN indoor localization", *Digital Communications and Networks*, Vol. 5, No. 4, 2019, doi: 10.1016/j.dcan.2019.09.001.

[15] H. Yao, D. Fu, P. Zhang, M. Li, and Y. Liu, "MSML: A novel multilevel semi-supervised machine learning framework for intrusion detection system", *IEEE Internet Things J*, Vol. 6, No. 2, pp. 1949–1959, 2019, doi: 10.1109/JIOT.2018.2873125.

[16] H. Chen, C. Meng, Z. Shan, Z. Fu, and B. K. Bhargava, "A novel low-rate denial of service attack detection approach in zigbee wireless sensor network by combining hilbert-huang transformation and trust evaluation", *IEEE Access*, Vol. 7, pp. 32853–32866, 2019, doi: 10.1109/ACCESS.2019.2903816.

[17] J. Hu, D. Ma, C. Liu, Z. Shi, H. Yan, and C. Hu, "Network Security Situation Prediction Based on MR-SVM", *IEEE Access*, Vol. 7, 2019, doi: 10.1109/ACCESS.2019.2939490.

[18] J. Liu, B. Kantarci, and C. Adams, "Machine learning-driven intrusion detection for Contiki-NG-based IoT networks exposed to NSL-KDD dataset", In: *WiseML 2020 - Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020. doi: 10.1145/3395352.3402621.

[19] D. Hemanand, D. S. Jayalakshmi, U. Ghosh, A. Balasundaram, P. Vijayakumar, and P. K. Sharma, "Enabling Sustainable Energy for Smart Environment Using 5G Wireless Communication and Internet of Things", *IEEE Wirel Commun*, Vol. 28, No. 6, 2021, doi: 10.1109/MWC.013.2100158.

[20] D. S. Jayalakshmi, D. Hemanand, G. M. Kumar, and M. M. Rani, "An efficient route failure detection mechanism with energy efficient routing (Eer) protocol in manet", *International Journal of Computer Network and Information Security*, Vol. 13, No. 2, 2021, doi: 10.5815/IJCNIS.2021.02.02.

[21] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "Mqttset, a new dataset for machine learning techniques on mqtt", *Sensors (Switzerland)*, Vol. 20, No. 22, pp. 1–17, Nov. 2020, doi: 10.3390/s20226578.

[22] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset", *Cluster Comput*, Vol. 23, No. 2, pp. 1397–1418, Jun. 2020, doi: 10.1007/s10586-019-03008-x.

[23] J. Makhija, A. A. Shetty, and A. Bangera, "Classification of Attacks on MQTT-Based IoT System Using Machine Learning Techniques", 2022, doi: 10.1007/978-981-16-3071-2_19.

[24] D. Hemanand, G. V. Reddy, S. S. Babu, K. R. Balmuri, T. Chitra, and S. Gopalakrishnan, "An Intelligent Intrusion Detection and Classification System Using CSGO-LSVM Model for Wireless Sensor Networks (WSNs)", *International Journal of Intelligent Systems and Applications in Engineering*, Vol. 10 No. 3. Ismail Saritas: 285–93, 2022.

[25] O. Almomani, "A feature selection model for network intrusion detection system based on pso, gwo, ffa and ga algorithms", *Symmetry (Basel)*, Vol. 12, No. 6, pp. 1–20, 2020, doi: 10.3390/sym12061046.

[26] S. S. Wali and M. N. Abdullah, "Efficient energy for one node and multi-nodes of wireless body area network", *International Journal of Electrical and Computer Engineering*, Vol. 12, No. 1, pp. 914-923, 2022, doi: 10.11591/ijece.v12i1.

[27] K. Cho and Y. Cho, "Hyper ledger fabric-based proactive defense against inside attackers in the WSN with trust mechanism", *Electronics (Switzerland)*, Vol. 9, No. 10, 2020, doi: 10.3390/electronics9101659.

[28] A. Ghosal and S. Halder, "A survey on energy efficient intrusion detection in wireless sensor networks", *J Ambient Intell Smart Environ*, Vol. 9, No. 2, 2017, doi: 10.3233/AIS-170426.

[29] N. Jameel and H. S. Abdullah, "Intelligent Feature Selection Methods: A Survey", *Engineering and Technology Journal*, Vol. 39, No. 1B, 2021, doi: 10.30684/etj.v39i1b.1623.

[30] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection", *Machine Intelligence and Pattern Recognition*, 1994, doi: 10.1016/B978-0-444-81892-8.50040-7.

[31] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", *Computers and Electrical Engineering*, Vol. 40, No. 1, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[32] K. Guolin, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, pp. 3147–55, 2017.

[33] H. M. Fadhil, M. N. Abdullah, and M. I. Younis, "A Framework for Predicting Airfare Prices Using Machine Learning", *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, Vol. 22, No. 3, 2022, doi: 10.33103/uot.ijccce.22.3.8.

[34] Q. Li, C. Tai, and E. Weinan, "Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations", *Journal of Machine Learning Research*, Vol. 20, 2019.

[35] I. Almomani, B. A. Kasasbeh, and M. A. Akhras, "WSN-DS: A Dataset for Intrusion Detection Systems in Wireless Sensor Networks", *J Sens*, Vol. 2016, 2016, doi: 10.1155/2016/4731953.

[36] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)", In: *2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings*, 2015, doi: 10.1109/MilCIS.2015.7348942.

[37] Q. Liu, D. Wang, Y. Jia, S. Luo, and C. Wang, "A multi-task based deep learning approach for intrusion detection", *Knowl Based Syst*, Vol. 238, 2022, doi: 10.1016/j.knosys.2021.107852.

[38] J. Su, Y. Chen, Z. Sheng, Z. Huang, and A. X. Liu, "From M-Ary Query to Bit Query: A New Strategy for Efficient Large-Scale RFID Identification", *IEEE Transactions on Communications*, Vol. 68, No. 4, 2020, doi: 10.1109/TCOMM.2020.2968438.

[39] T. S. Sabbah, "Hybridized Dimensionality Reduction Method for Machine Learning based Web Pages Classification", *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, Vol. 22, No. 3, 2022, doi: 10.33103/uot.ijccce.22.3.9.

[40] I. M. Bapiyev, B. H. Aitchanov, I. A. Tereikovskyi, L. A. Tereikovska, and A. A. Korchenko, "Deep neural networks in cyber attack detection systems", *International Journal of Civil Engineering and Technology*, Vol. 8, No. 11, pp. 1086–1092, 2017.

[41] S. Mohammadi, H. Mirvaziri, M. G. Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm", *Journal of Information Security and Applications*, Vol. 44, pp. 80–88, 2019, doi: 10.1016/j.jisa.2018.11.007.

[42] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network Intrusion Detection Combined Hybrid Sampling with Deep Hierarchical Network", *IEEE Access*, Vol. 8, 2020, doi: 10.1109/ACCESS.2020.2973730.