International Journal of Mathematics, Statistics, and Computer Science
Vol. 2
ISSN: 2704-1077 eISSN 2704-1069, DOI: https://doi.org/10.59543/ijmscs.v2i.10377

Enhancing Intrusion Detection in Wireless Sensor Networks through Machine Learning Techniques and Context Awareness Integration

Osamah Ahmed Mahmood[1]
[1]Department of Computer Science, Modern University for Business and Science, Damour, Lebanon; oamdhk32@gmail.com

Received 11.05.2024, Revised 24.06.2024, Accepted 09.07.2024, Published 02.08.2024

ABSTRACT Wireless sensor networks (WSNs) play an imperative role in the communication among independently implemented wired, localized, or mobile sensors in a cyber-physical system. Environmental monitoring, object identification, data acquisition, analysis, and transmission to the owner of the wireless sensor network are the primary research focuses in this work. Due to the flexibility of WSNs and the scarcity of resources in IoTs, both are frequently integrated. But the above integration makes these networks more open to outside factors and, thus, vulnerable to attacks like flooding, assaults, grayhole attacks, black hole attacks, and other scheduling problems that are typical in such circumstances. There are certain inherent properties of Wireless Sensor Networks (WSNs) that have made the detection of intrusion unsuitable, namely, false alarms, it has very high computing overhead and a poor detection ratio. This problem has arisen due to the network's excess of data tightly connected in a dense manner and resource constraints of sensor nodes. For that purpose, this research recommends the utilization of machine learning techniques for intrusion detection in WSNs. The detection precision is enhanced by enhancing the use of Support Vector Machine (SVM) combined with stochastic gradient descent (SGD). Moreover, the research proposes the integration of context knowledge, which is known as context awareness that takes into account user preferences and system characteristics or situations to improve the performance of the recommendation systems. In order to decentralize the computational load of the system, the first traffic data is reduced by principal component analysis (PCA) and singular value decomposition (SVD). The network risks observed are further categorised using an VG-IDS model. Worthy of note is the fact that the recommended WSN-DS algorithm could yield better results than other complex algorithms that were examined by using the WSN-DS dataset throughout the evaluation process; the accuracy rate of the proposed WSN-DS algorithm was 96%. It is seen in improvement of accuracy and recall and F1-measure rates to the improved figure of 98%, 96%, and 97% respectively.

Keywords: Intrusion Detection; Internet of Things; Wireless Sensor Network; Accuracy; Machine Learning.

1. INTRODUCTION

As a result, of Wireless Sensor Networks WSNs, The Internet of Things (IoT) is directly attributed to wireless sensor networks. . This is well explained with the help of Internet of Things, which is able to connect two or more gadgets and significantly change people's lives [1-2]. It is important to state that to provide the highest level of security to an IoT network, it is required to deploy certain specific measures. WSN has adopted various security measures such as Authentication, Encryption and all other related ideas into WSN. On the other hand, there are new security threats that can potentially bypass conventional solutions since there is diversity in attack techniques. Therefore, the confidentiality of information is the main condition that should be given special attention when creating a WSN to connect a large number of devices. Security of WSN systems must be maintained and enhanced; as this is the case, security in WSN system is becoming stricter in stringency. [3-4].

However, passive defensive strategies are not enough to provide Wireless Sensor Networks (WSNs) utterly security solutions. The works hold that there should be [5] preventive safety technologies on board. It is wise to use an IDS for constructing active defense configurations [6]. When conventional preventive mechanisms are not possible, then the utilization of Intrusion Detection Systems (IDS) which incorporates data-oriented methods can be employed to actively detect hostile incursions. The volume of the traffic, which a network transmits, rises, and that is why IDS have problems with judging it in real-time manner. The behaviour of data in WSN is another factor that has to do with the rate at which data has to be processed and hence defines the effectiveness of IDS [7].

Therefore, the first sign of an abnormal traffic can be signalled by sudden emergence of parameter features in the WSN, and the advent of traffic is relatively quicker. The interconnection normalcy can be further affected by calamities including flooding, jamming, sink holes, and worm hole [8]. Because of huge quantity and variety of data and absence of data, the classifier can have issues with fast identification of normal and extranormal traffic patterns [9]. The problem is how to identify the presence of malicious traffic in the network

among other non-related information flow and other extra elements. This issue extends the hours and energy used in searching for answers while at the same time reducing the chances of getting them [10].

The dataset and features extracted utilized in the study or method determine machine learning's limits in the main. This highlights how important feature selection is to machine learning. Databases with tens of thousands of dimensions in feature space are becoming more common as computers are used in more areas of daily life. Nevertheless, only a small set of characteristics are able to fully capture the image. This data subset has a large number of redundant and unnecessary features, which greatly degrades machine learning system performance. The scientific community has demonstrated that the combination of feature selection with machine learning is quite successful, and this has resulted in the creation of technology that is extensively used in domains like networks traffic observing and safety [11].

Reducing the time, it takes to identify intrusions without compromising their accuracy or detection rate should be a top priority. Conventional approaches to network intrusion detection are insufficient for protecting WSNs due to the inherent incompatibilities among WSNs and conventional computer systems. Various aspects, including network design, data transfer, terminal types, and more, differ. Above all else, a WSN IDS must accurately and dependably detect both well-known and unidentified security threats to a wireless sensor network. Furthermore, the lightweight intrusion detection system (IDS) for Wireless Sensor Networks (WSNs) is essential for preventing a substantial decrease in the infrastructure's functionality. [11]. This paper provides a new perspective which is perfectly suitable for identifying anomalies in Wireless Sensor Networks (WSNs). This work's primary contributions are:

- The computational costly intrusion detection method and low recognition rate of intrusion activities are both caused by the massive amounts and diverse types of data that the WSN must process. Consequently, in this intrusion detection study conducted on WSNs, there has been the consideration of the following feature selection methods: SVD and PCA.
- Thus, the research has the purpose of introducing a new intrusion detection model VG-IDS through examining and comparing two different categorization algorithms. Stochastic Gradient Descent and Guessing Naive Bayes in the WSN problem introduction.
- In WSN, the VG-IDS is used to detect traffic attacks with a lower frequency of false positives. The shortcomings of conventional Wireless Sensor Network (WSN) intrusion detection techniques are attempted to be addressed by this paradigm. These drawbacks include poor detection performance, sluggish real-time performance, and out-of-date findings in comparison to related studies.

The following overview describes the structure of the remaining portions of this work. The detection of intrusions in wireless sensor networks (WSN) is examined in Section 2. Relevant study findings are presented in Section 3. Section 4 reveals the strategy for preventing unwanted access to wireless sensor networks. In Section 5, experimental habitats are shown. Section 6 contains the outcomes and examination of the experimentations. Section 7 outlines the objectives for the future.

2. DETECTION OF INCURSIONS IN WSN

In the domain of Wireless Sensor Networks (WSNs), attacks are categorized as either passive or hostile. Also known as destructive attacks, active attacks are immediate and direct threats to the system that are perpetrated by adversaries. Conversely, passive attacks involve the source station taking over the data that is useful to the destination station while not interrupting the typical data flow. Individuals who are not authorized can gain access to data that is legitimate, this can lead to significant issues with the networks and lead to numbers of safety concerns. The issue of sensitive material does not impede its transmission. Passive attacks, such as eavesdropping on the conversation of a node, are different from active attacks, which take advantage of the broadcast nature of wireless communication [12]. The nature of the attack, whether internal or external, is influenced by its primary cause. To steal personal information from a Wireless Sensor Network (WSN), criminals only need to have access to powerful wireless devices for sending and receiving signals. These assaults often utilize methods like injection, replay, eavesdropping, and interference. The loss of a critical network component can facilitate an attack from a domestic threat. Within internal attacks on a network, there are two types of nodes: legitimate nodes and malicious observers. Independent nodes don't rely on other nodes and only use network resources without directly harming other nodes [13]. Conversely, malicious sensor nodes embody by imitating typical nodes and participating in actions like eavesdropping, interrupting, or altering the way communication is conducted across the entire network. Communication bandwidth, processing power, Energy and memory available are all finite resources for wireless sensor networks (WSNs). Therefore,

every situation and setting have unique needs, and these needs dictate how intrusion detection systems are designed.

3. RELATED WORKS

Wireless sensor networks (WSNs) and ad hoc networks, which are both mobile and wireless, are examples of bigger wireless systems that are unable to be adequately protected by conventional wired intrusion detection systems (IDS). To resolve this issue, it is crucial to include intrusion detection technologies into wireless sensor networks (WSN). Anomaly detection-based intrusion detection systems are those that rely on unusual activity to guide their design. In view of the said observations, researchers have come up with a number of anomalous detection methods. These group of approaches owe their origin to cluster analysis, statistical learning models, machine learning algorithms and artificial immunity methods.

To avert the challenges associated with the intrusion identification models, a multilevel semisupervised ML (MSML) architecture was developed by Yao et al. [15]. Therefore, the design of the present study comprises four different methods, namely pattern recognition, FC, model modification, and the extraction of discrete clusters of points. Employing a hierarchical clustering method, the most important component of the semisupervised approach, namely the determination of "pure clusters," is completed.

In a WSN a study by Chen et al. [16] offered a method of identifying LDoS assaults by using the HHT. While it may be impractical to aim at eliminating the suspects entirely, it would be possible to locate a considerable number of reliable nodes at a high rank which, if merged, would form a good starting point in producing the culprits. Reduction of energy utilization, time, and traffic were the goals to be achieved.

Hu et al. [17] improved WSN security in a separate study by merging the SVM classification approach with the CSO algorithm. Finding a realistic and efficient method to use the existing data to forecast network incursions was the driving force behind this study. The SVM classification model achieved its objective of spreading the parameters among the nodes by employing the map reduction method. Nevertheless, one must take into account the following drawbacks: a) Participant results are missing, b) reaction time is slow, and c) overclassification is common.

To identify novel NSL-KDD archive data, Liu et al. [18] used EM in their study. In this post, we looked at several types of attacks, including synflood, land, ping of death, sweeping, and UDP flood. In order to optimise energy consumption and resource distribution, Hemanand et al. [19] suggested using the Glow-worm Method with Internet of Things (IoT) sensors. This approach enables eco-conscious, intelligent energy management. It is impossible to assess a network's efficiency without taking its routing protocol into account. To make the network more secure, Jayalakshmi et al. [20] proposed encrypting all of the nodes. When optimisation methods such as PSO, GWO, FFA, and GA are available, Almomani argues that NIDSs can be made more effective by merging feature selection models. The suggested paradigm uses thirteen rules extracted from the aforementioned algorithms in an effort to boost NIDS's performance. To finish the aforementioned implementations, Anaconda's Python Open Source and wrapper-based techniques were crucial. Using the UNSW-NB15 dataset with the SVM and J48 classifiers, we evaluated the suggested model's performance.

The MQTT protocol is widely used by IoT devices, and Chang et al. [21] designed a dataset named MQTTset with that protocol in mind. The fact that academics have developed a mechanism to identify fraudulent actions in a real dataset—one that include security assaults on MQTT networks—is further proof that the dataset is legitimate. So, the findings of this research are the demonstration of the potential of MQTTset in developing machine learning models to increase security in IOT networks.

In their research, Kumar et al. [22] proposed a technique of intrusion detection by using the concept of violating usage. This method has the capability of distinguishing different types of network assaults such as probes, assaults that bring about denial of services, often special probes alongside normal attacks. The authors marked the utilization of the KDD or NSL-KDD 99 dataset in projects linked with Intrusion detection Systems (IDS). Although, it is now more or less realised that such records are archaic and ineffective to establish modern threats. To counter this problem, the research team in this study leveraged on an approach involving the UNSW-NB15 dataset that contains real world features not in the database so as to build a model that would be able to distinguish cyber crime from normal traffic. Chandre et al. [23] also showed in another research the potential of different AI models in predicting the level of attack success on the IoT systems using MQTT. The asses focused on the models taking the parameters such as precision, accuracy, and F1 scores into consideration. The study thus revealed that the outcome of the random forest method was more effective than the other models for having a near perfect estimated accuracy rate of about 96 percent.

Hemanand et al proposed a new IDS for WSNs; they called it "A hybrid IDS that utilises LSVM and CSGO" to enhance the security of WSNs. This system integrates aspects of one known as Cuckoo Search Greedy Optimisation (CSGO) and the other known as Linear Support Vector Machines (LSVM). Experiments are carried out with NSL-KDD and UNSW-NB15 datasets commonly used in the network field to evaluate the

performance of this IDS method. The pre-processing of the data involves attribute smoothing those entails employing CART for categorization.

This requires data duplication, missing value, and appropriate criteria for data selection and management. According to the ideal number of features determined during pre-processing, the CSGO method is used here. Finally, the LSVM machine learning method is utilized for a determination of the label as either normal or anomalous. Therefore, it is possible to confirm how well the suggested security architecture addresses the organisational performance indicators by using several measures.

In their original study, Salmi and Oughdir [25] applied a number of tests created by employing the NS-2 network simulator that adopted the LEACH protocol. The objective of these exercises was to accumulate a great deal of network data which were then subsequently altered in order to obtain 23 various network parameters which served as a proxy for the status of the relevant sensor. Five forms of DoS attacks were also used throughout the study; they included Results for the 25-epoch CNN-LSTM model evaluation was as follows: 944 for accuracy, 0 **mean or average**. 96 for precision, and 0. 922 for recall. The contribution of these findings to the research was given a value between zero and one.

4. WSN INTRUSION DETECTION ARCHITECTURE

The Wireless Sensor Network (WSN) intrusion detection system consists of three main components: There exists a data collection facility, IDS model, and a reaction facility known as a reaction module that deals with intrusions discerned. Procedural sub-module of Data Collection Module involved in the gathering of information from the environment while the sub-module of the Detection Module involves the processing of the gathered information. In this particular module, analysts are used to analyze and assess the collected data with the primary objective of determining instances of intrusion in WSN. If an irregularity is detected, then the detection module is to notify the reaction module of this right away. Figure (1) show the application of WSN for intrusion detection area, it shows the sensor nodes (SN), cluster head (CH) and sink nodes (Sink) that forms the WSN. [26, 27].

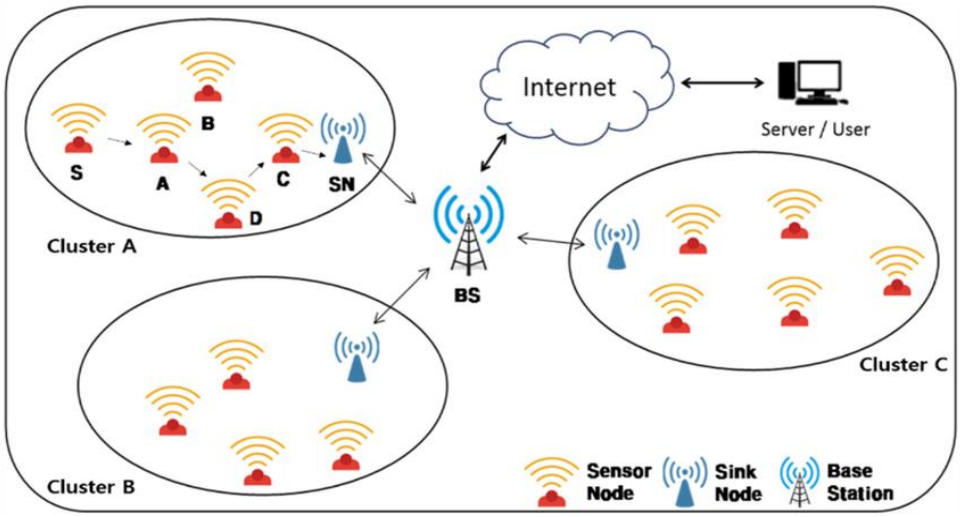

Figure 1. Initial work for intrusion detection for WSN by using a basic design [28].

This distributed detection system for intrusion reduces the volume of communication and energy expenditure. The primary task of the cluster head is to oversee and regulate the various computer processes that occur globally. By reducing the volume of data to be transmitted and having only the cluster head communicate this volume, the system achieves effective communication. Additionally, the cluster head can task regular sensors with the computational burden, thereby alleviating the load. To that end, scientists have experimented with more complex data mining techniques in an effort to make IDSs in WSNs more effective. Due to the large computational costs, however, these solutions are impractical for real-time systems that utilise WSNs. huge input data dimensions, duplicate data in huge datasets, and inadequate data pre-processor are the main reasons for the great costly expenditure of IDS.

4. THE PROPOSED IDS APPROACH

Reducing the number of potential features is the main objective of feature selection for managing their quantity. The goals here are to preserve as much data as possible, make the IDS's computations easier, and improve the classification algorithm's accuracy. This is achieved by utilising commonly used pre-processor

techniques, like principal component analysis and singular value decomposition. There is a data analysis part of the programme that uses these techniques to check if the behaviour is suspicious. The suggested architectural design is described in full in Figure 2. Phase one of the study entails looking at various ways to manipulate the data in this particular setting.

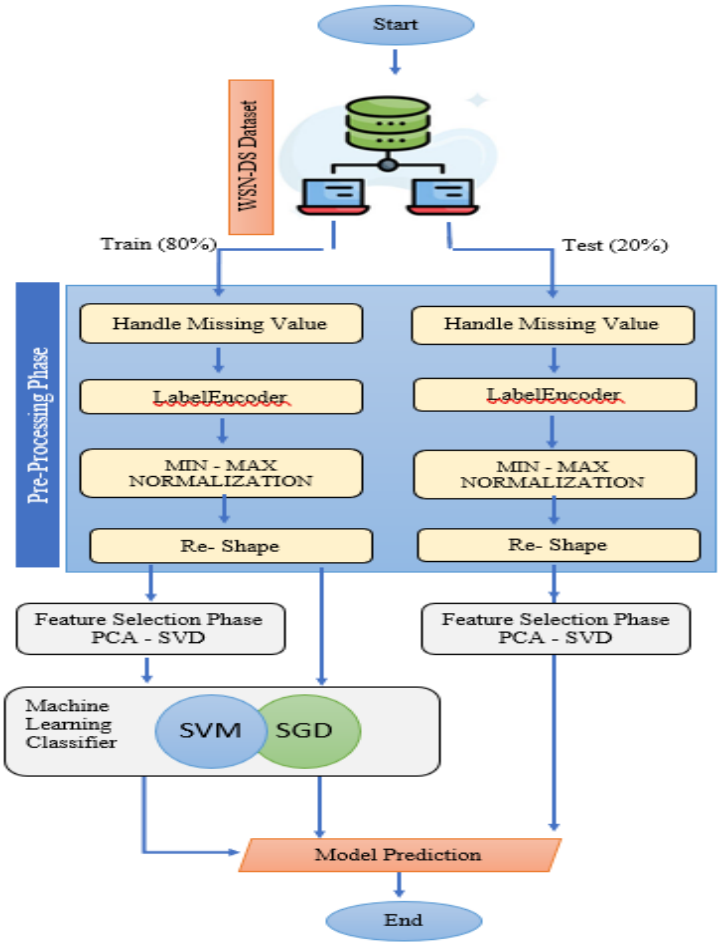

Figure 2. The VG-IDS algorithm framework specification.

Wireless sensor networks use a combination of machine learning techniques to identify intrusions as efficiently as possible. First, the input data must be filled in. Then, a sequence of operations must be carried out within the constraints of certain parameters and hyperparameters. The dimensionality of the dataset is reduced using Principal Component Analysis (PCA), which retains the most relevant properties. Furthermore, two distinct datasets are created for training and testing; this is accomplished by using Singular Value Decomposition (SVD) to improve the data and draw attention to key features. The allocation is 80/20, meaning that testing receives 20% of the overall budget and training receives 80%. At this stage, we use two models: a Support Vector Machine (SVM) and a Stochastic Gradient Descent (SGD). However, SVM also maintains the mean and covariance of each class distinctly and SVM also calculates the probability of each class. Conventional SG then is trained from data which is raw or in other words, is not restricted by the training or the SGD technique. Therefore, this algorithm also estimates the gradient of said loss function with respect to the model's parameters.

This step in the data engineering stage that is charged with the evaluation of data significantly influences the performance of well-developed models. The PCA and SVD utilize the transformations in a manner that ensures it can be used to forecast the test set individual points. For making a prediction, the SVM model uses merely a class probability and then identifies the probability range which pertains to the concerned class. In predicting, SGD's attempt to estimate parameters that place the decision boundary is done using stochastic gradient descent. By calculating the F1 score, recall, accuracy, and precision for both anticipated and actual labels, we can evaluate the system's performance, a strategy known as "data engineering" that is fundamental to learning to detect invasions.Cleaning, normalising, and choosing characteristics are the three distinct phases of data processing. Using a filter-based approach that incorporates principal component analysis and singular value decomposition, the most important features are extracted. As soon as the correct feature

vector is identified, the model is assigned to the training set. The model's efficacy is evaluated when training is finished by comparing its results to the validation set. In the end, the experimental dataset is analysed using the validated model.

Algorithm (1): The VG-IDS aproach steps

Input: load the dataset called WSN-DS.
Output: Intrusion Classification
Start:
Define Parameters:
Specify the value of the number of components for PCA, which is num_components.
- **Tune the parameters which are specific to SVM, like the type of the Kernel function to be used and the value of C, the regularization parameter.**
- **Specify the choices of hyperparameters as follows for SGD (learning rate and number of iteration).**

Initialize Storage:
Make lists to store true labels and the respective predicting labels.
- **Dimensionality Reduction:**
- **PCA needs to be used to select num_components of features and decrease the dimensionality of the dataset.**
- **By performing D, the number of features to be considered is still large. Application of SVD can be used to pre-process them and select the most appropriate ones.**

Split the Dataset:
Split the data set into raw train (80%) and raw test (20) set.
Train the Models:
Training SVM (Support Vector Machine):Training SVM (Support Vector Machine):
- **Create the SVM model with the specific parameters that were set in this section.**
- **Pass the training dataset to the model for it to learn through an appropriate algorithm (for this a Support Vector Machines, SVM model is often used).**

Training SGD (Stochastic Gradient Descent):
For each iteration:
- **In this and the next step we can sort the training data, and then turn on the random selection of data, randomly shuffle the training data.**
- **For each data point (x, y) in the training data:**
- **Find the derivative of the loss function with respect to all the train parameters of the model.**
- **Update model parameters using the formula: Here, we have the following adjustments: parameters = parameters – learning_rate * gradient.**

Test the Trained Models:
For each data point (x_test, y_test) in the testing set:For each data point (x_test, y_test) in the testing set:
- **Must complement transformation assignments with the help of PCA and SVD techniques.**

Predictions Using SVM:
- **Based on the trained SVM model you should be able to predict the class label of the data point.**

Predictions Using SGD:
- **Compute the decision boundary and use the parameters learned by using SGD.**

Put the data point into the appropriate categories according to this decision.
Evaluate Performance:
For each predicted label and true label:
- **Compute evaluation metrics: Sensitivity, Specificity, Area Under Curve and F1 statistic.**
- **Print the evaluation metrics.**

End

4.1. The Pre-processing for Dataset

4.1.1. Gathering and representing data

The delivered data is a property with alphabetic letters that, to circumvent it in the method, must be transformed into numerical values. The classification of attacks has five different types: "Normal," "Blackhole," "Grayhole," "Flooding," and "TDMA." Because it's impossible to quantify this information, the data are organized by ordinal numbers from 0 to 4 in a logical manner. For any changes or revisions that may be required, refer to Table (1). [30].

Table 1. Characteristic values and their conversions across attacks.

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

Flooding 4

4.1.2. Label encoder

The feature names, regardless of their nominal or ordinal nature, are composed and represented as strings. Some labels may necessitate the organization of information based on the order of attributes, while others may not require the organization of information based on nominal attributes. Converting labels into numerical values during the pre-processing of data is crucial to ensuring the learning algorithm is comprehended by the features appropriately. The LabelEncoder employs a numerical method of encoding to associate data with labels.

4.1.3. maximum and minimum normalization

Continuous data normalisation is required in many classification methods that are significantly impacted by the magnitude of data characteristics varying from small values to those in hundreds of thousands. Here, we take the extreme values obtained from Equation (1) as a benchmark point for our analysis. The initial data for the jth feature dimension is denoted as xj, where Minj and Maxj represent the minimum and maximum value of the feature, respectively. The normalized data for the feature is also denoted as xj [30].

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

4.2. Features Extraction

4.2.1. Principal Component Analysis

Pattern recognition in high-dimensional space is a common problem that principal component analysis (PCA) attempts to solve. Using a modest set of distinguishing features, principal component analysis (PCA) aims to describe images. This set of images can represent both known and unknown faces. Through PCA, we can identify statistical evidence that supports the identification of facial traits, this information is then used to strengthen the legitimacy of Principal Component Analysis. To utilize PCA, we must convert a 2D matrix of facial images into a single dimension; it is important to recognize that a single dimension can be represented in either horizontal or vertical form, without altering its nature. singular value decomposition [22,23].

4.2.2. Singular value decomposition

Singular value decomposition (SVD) is a different method of data division. It has multiple uses in signal processing and statistics, one of these is the identification of patterns and the extraction of features from vectors. On the other hand, PCA can't figure out what features are in a signal that shows frequency variations, and it can't figure out what features are in a signal from only one signal. Singular Value Decomposition (SVD) is more beneficial than PCA because it can differentiate between the states of the body that are actually present, but which are masked by the different frequencies that they have [31-33].

4. 3. Classification Models

Intrusion detection is accomplished by utilizing the VG-IDS algorithm for classification with preprocessed data from wireless sensors, and then employing a method called sequence backward feature selection. VG-IDS is a system that employs gradient-based approaches. It's renowned for its quick processing, decentralized nature, and superior performance. The core principle of VG-IDS is a training method that utilizes a histogram-based approach, this method is limited in its use of features and samples.

4.3.1. Support Vector Machine

Another well-liked machine learning technique, Support Vector Machine (SVM) is mainly employed for regression and classification. When high dimensionality or more dimensions than samples are present, support vector machines (SVMs) are thought to work best. The main goal of support vector machine (SVM) is to locate the optimal hyperplane in a space with N dimensions, where N is the number of qualities that distinguish one instance from another in the data.

Choosing the optimal hyperplane for class differentiation is the key component of the support vector machine algorithm. A hyperplane is considered ideal if it maximises the distance from the nearest point in both classes. In the training set, you'll find these data points that are closest to the class that will be separated from another class; they're called support vectors as shown in figure 3[34].

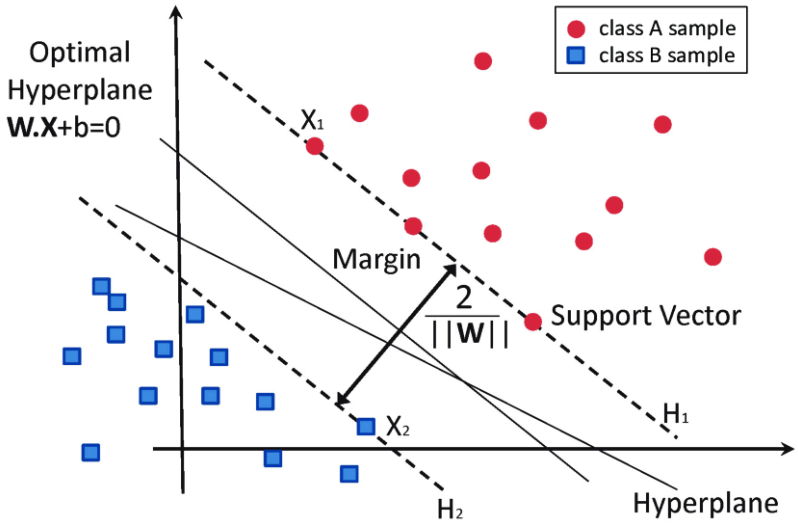

Figure 3. A graphical representation of SVM algorithm classifier.

4.3.2. Stochastic Gradient Descent

For linear classifiers in particular, the model is an outstanding learning strategy. It is sufficient to deduct the real gradient from the predicted one. With stochastic gradient descent, we can estimate the function's gradient by calculating a gradient for each learning component linked to the cost function. In order to accomplish the necessary changes, the settings were adjusted multiple times. Whenever new training data was collected, the model's parameters were adjusted. Stochastic gradient descent outperforms the traditional approach when used to big datasets [35]. This method of facilitating is really working. All things considered, the updated stochastic gradient descent (SGD) equations in (2) read as:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_t \nabla l_i\left(\theta^{(t)}\right) \qquad (2)$$

When adjusting the settings, both variables reflect the size of the learning set. You can see the exact number of iterations in the variable 't'. Index I will be arbitrarily assigned a new value before each repeat. Randomising samples before analysis is a common practice, nevertheless [36].

5. EXPERIMENT CONFIGURATION

The dataset used in this study is public and can anyone have used [37]. This dataset includes metrics that could be utilised to detect possible hacks, and it is designed to be interoperable with Wireless Sensor Networks (WSNs). The four most common types of Denial of Service (DoS) assaults in the context of Wireless Sensor Networks with Data Sink (WSN-DS) are blackholes, grayholes, floods, and scheduling attacks. The specific details are listed in Table 2. The training set was composed of 299,728 samples that were randomly selected, which represents around 80% of the total. Conversely, the testing set was composed of 74,932 samples that were randomly picked, which made up the remaining 20%.

Table 2. The WSN-DS dataset's class labels description.

Class	Description
Normal	Logs of a Typical Link
Blackhole	When an attacker initially identifies himself as a CH, they are launching a denial-of-service assault against the LEACH protocol.
Grayhole	The LEACH protocol is the target of a denial-of-service attack, which begins with the attacker posing as a CH to other nodes.
Flooding	There are multiple entry points for an attacker to compromise the LEACH protocol.
Scheduling	While LEACH is being initialised, its scheduling infrastructure is vulnerable to attacks.

The F-measure, recall, precision, and accuracy of the dataset are evaluated using the confusion matrix (CM). Using the method described in equations (3) to (6), which are cited in references [38,39], we can find the optimal balance between the number of false positive results and the ratio of false negative results to the total number of outcomes. The percentage of properly annotated training data is called precision. Try again with the percentage of "good" things that ended up in the "good" pile. The accuracy of a detection model is measured by the number of false positives it produces, which occur when certain things are incorrectly labelled as negative regardless of their true classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

6. RESULTS ANALYSIS AND DISCUSSIONS

This study, which comprised multiple separate experiments, proved that the proposed concept worked. The main goal of this experiment is to compare the performance of VG-IDS using PCA/SVD and 10 or 15 features, with and without a feature extraction phase. Examining how VG-IDS stacks up against competing machine-learning categorization methods is the secondary goal. In this study, we will test VG-IDS against four different sorts of attacks to see how well it performs. Achieving a high level of performance in detecting potential threats is essential for the smooth integration and proper functioning of the network's intrusion detection system. In order to assess this performance, the precision and recall metrics are of utmost importance. Table (3) shows the results of a comparison between the performance of DNN [21] and Deep CNN [23] on the WSN-DS dataset without feature selection and that of popular classification methods like Stochastic Gradient Descent (SGD) and Gaussian Naive Bayes (SVM) without feature selection.

Table 3. Measure other classification metrics when feature selection is not applied, WSN-DS database was used.

For more accurate estimators or require obtaining a better fit on the higher order of dimensions for sample sets, you can use the feature selection module which is designed for feature selection or dimensionality reduction. Ten to fifteen features are used in the strategy of this work, and two methods: the PCA and the SVD. 4–7 are the plots illustrating the feature selection concerning the outcomes of applying machine learning methods to WSN-DS.

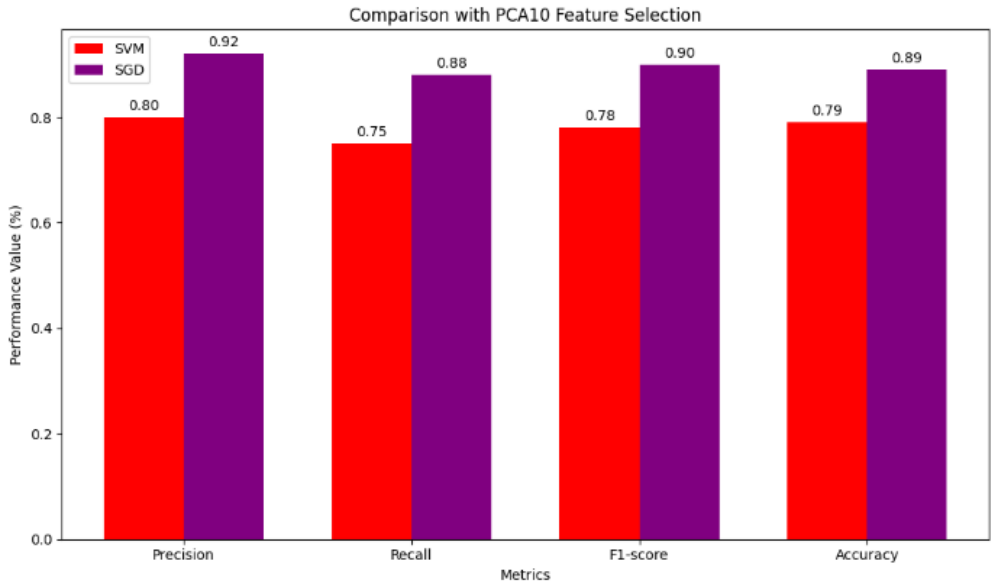

Figure 4. Comparative classification analysis of various feature selection metrics using technique (PCA10) on WSN-DS dataset.

Algorithms		Measures			
		Accuracy	Precision	Recall	F1-score
DNN [21]		0.98	0.93	0.87	0.90
Deep CNN [23]		97	94	92	90
SG-IDS	SVM	0.821	0.812	0.822	0.791
	SGD	0.952	0.981	0.953	0.961

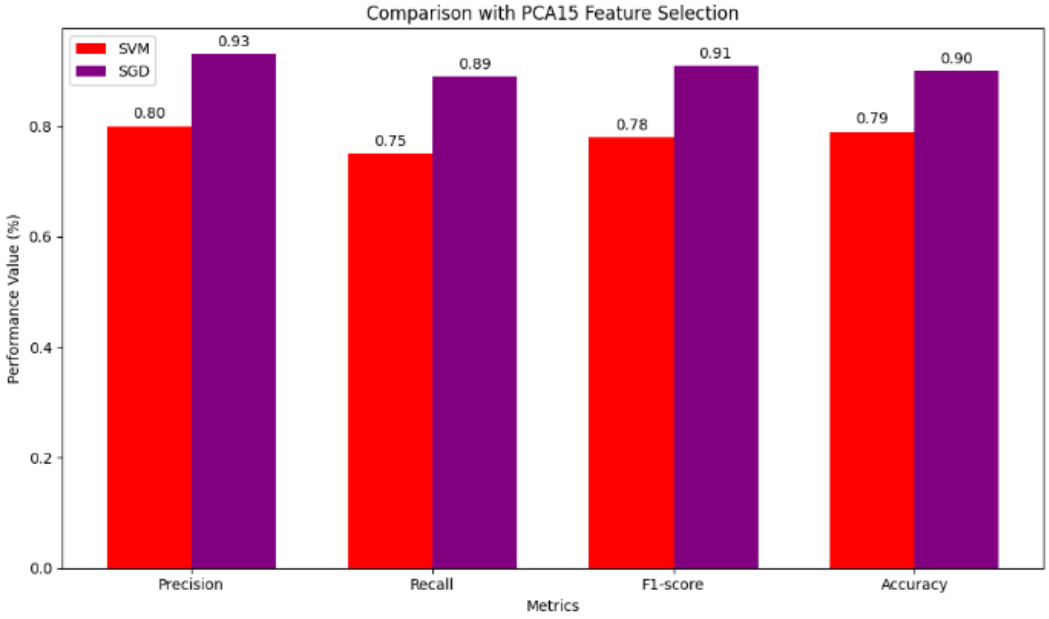

Figure 5. Comparative analysis of several classification metrics using feature selection (PCA15) on WSN-DS dataset.

Figure 6. Comparative analysis of several classification metrics using feature selection technique (SVD 10) on WSN-DS dataset.

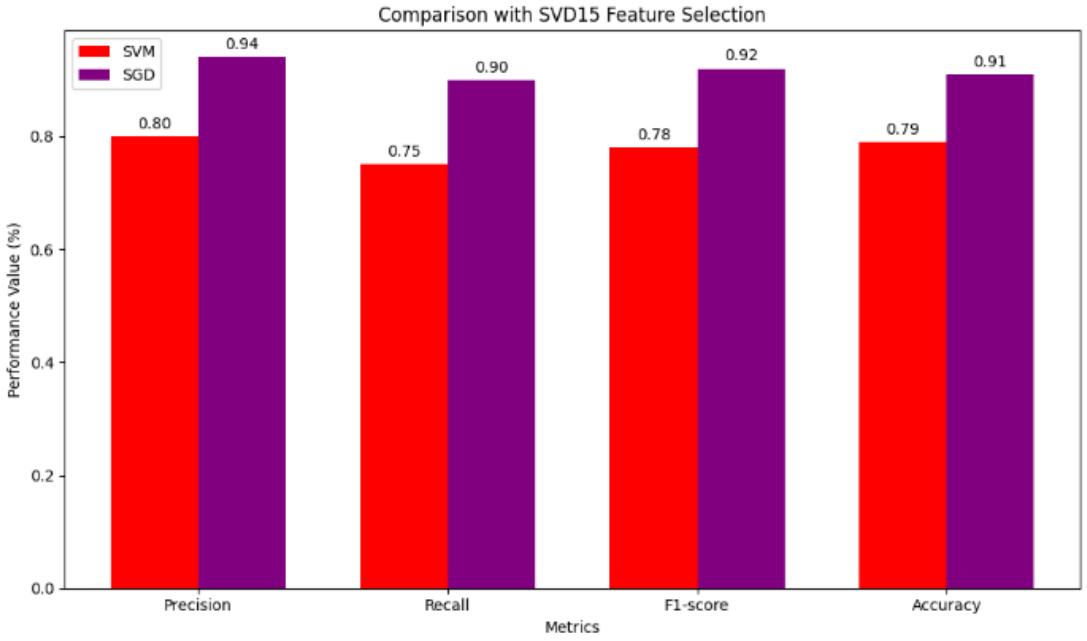

Figure 7. Comparative analysis of several classification metrics using feature selection technique (SVD 15) on WSN-DS dataset.

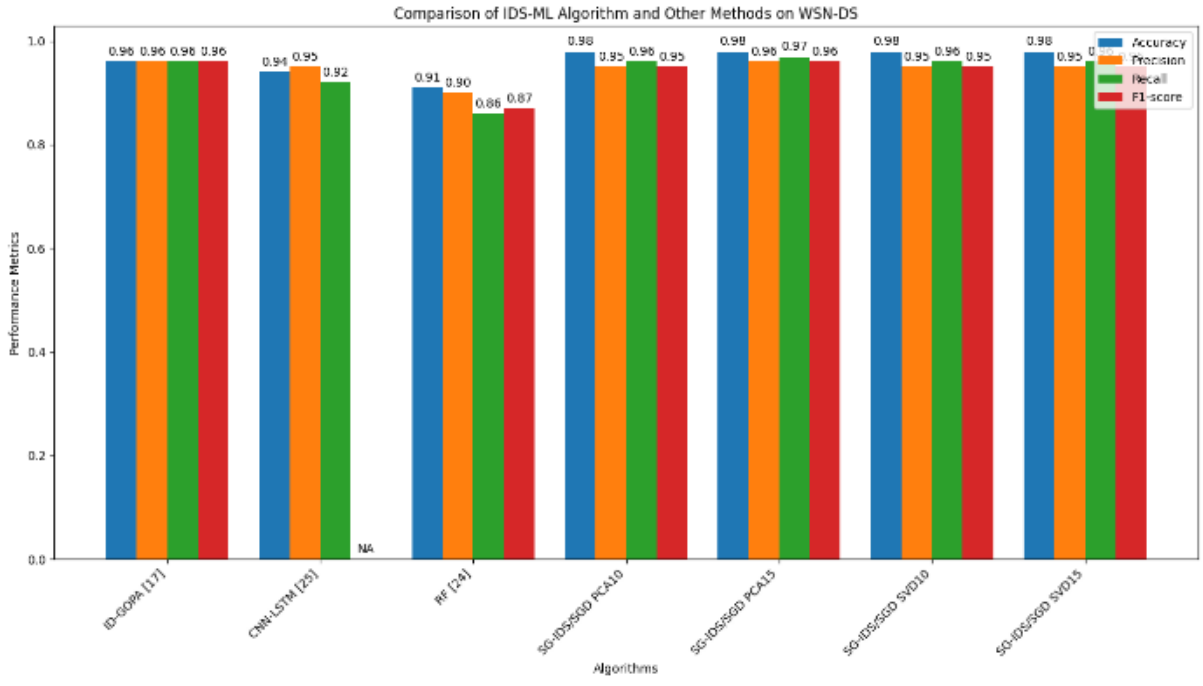

Figure 8. Analysing multiple classification measures.

The earlier investigation has yielded exact results, according to the scholars. Table (3) indicate that when used to the WSN-DS wireless sensor network dataset, the SGD algorithm outperforms other approaches in terms of accuracy and recall. It is feasible to modify the amplitude of the gradient according to the sample size using Stochastic Gradient Descent (SGD). A reduced gradient is the obvious consequence of a more precise model. The effect of feature selection on algorithm parameters such as accuracy and F-measure is

shown in Figures (4) to (8). Compared to previous approaches, the feature selection process performs better here. Skill in feature dependency analysis and knowledge of the connection between feature subset search and model selection are prerequisites for productive work with the WSN-DS dataset. Since the next three methods disregard the interplay of the classifier's data, eliminating superfluous intrinsic qualities is a breeze. Although these features have a lot of potential for difference, when taking the complete data into account, their capacity to differentiate is restricted. Considering the accuracy of predictions, the learning algorithm of the wrapper assesses the merits of a proposed subset. Classifiers and feature selection collaborate to find the most useful set of attributes for training. In contrast to other approaches, such as ID-GOPA [17], the VG-IDS algorithm outperforms RF [24] and CNN-LSTM [25] as shown in Table (4). This is because the sensor nodes use the traffic data, they acquire to initially choose the characteristics. A method that makes use of principal component analysis and singular value decomposition has been created to optimise the traffic characteristics. Minimising the quantity of traffic data while simultaneously increasing the model's accuracy are the primary goals of this method. But there are problems with feature selection and classification that prevent intrusion detection systems in wireless sensor networks from functioning in real-time. These problems are as follows: Low detection accuracy is one of these problems, and another one is related to system complexity. The approach discussed above as a prevention-based technique with the proved quality of the real-time performance and a powerful detection capability implies that each of these constraints is addressed separately. This approach successfully solves each of the above problem and also avoid the problem of overfitting.

7. CONCLUSIONS

Another aspect that is steadily being integrated into intrusion detection system is the selection of features to be used in conjunction with the machine learning techniques. Feature selection techniques aim at enhancing the model’s generalisation performance and minimizing model’s sensitivity and feature quantity. Also, these techniques can provide information about existing connections between attributes and their values. In order to compare the outcomes of different approaches commonly used in machine learning, the experiment was designed in an artificial setting. First, among the competitors, VG-IDS provides an outstanding accuracy due to the decision-based learning algorithm and gradient boosting architecture. Thus, it allows flexibility in its application for different tasks, efficiency in terms of using lower memory consumption and high accuracy ranging from 95-96 percent. Also, VG-IDS can accept a large amount of data sample for testing. Therefore, this research seeks to improve the usability of VG-IDS more by comparing and assessing various IDS algorithms for WSNs. This means that this evaluation will assist in realizing high intrusion detection rates without much computational intensity being felt. The aim of this approach is to decrease the number of features and remove all unimportant information, which, of course, removes any concern. Further, in the later part of the work, VG-IDS will be employed in the enhancement of accuracy and memory usage. As it was illustrated when comparing the real data with other similar systems and in the course of empirical research, such method is characterized by high indices of detection, low indices of false alarms, and low demands on the essential consumption of computational resources. This system has also the potential of detecting intruders within the wireless sensor networks securely.

Funding: This research received no external funding.
Conflict of interest: The authors declare no conflicts of interest.

References

[1] S. Pitafi, T. Anwar, I. D. M. Widia, and B. Yimwadsana, “Revolutionizing Perimeter Intrusion Detection: A Machine Learning-Driven Approach with Curated Dataset Generation for Enhanced Security,” IEEE Access, vol. 11, pp. 106954–106966, 2023, doi: 10.1109/access.2023.3318600.

[2] O. Chentoufi and K. Chougdali, “Intrusion Detection Systems based on Machine Learning,” Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning, 2021, doi: 10.5220/0010734300003101.

[3] D. Chatterjee, “An Efficient Intrusion Detection System on Various Datasets Using Machine Learning Techniques,” Machine Learning Techniques and Analytics for Cloud Security, pp. 103–128, Dec. 2021, doi: 10.1002/9781119764113.ch6.

[4] Joel Emmanuel Mulepa and Dr Glorindal Selvam, “Proficient Intrusion Detection System using Machine Learning using Machine Learning,” International Journal of Advanced Research in Science, Communication and Technology, pp. 499–506, Apr. 2023, doi: 10.48175/ijarsct-9072.

[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] S. P. Vijaya Vardan Reddy, S. P. Manonmani, C. Anitha, D. Jaganathan, R. Reena, and M. Suresh, “MLIDS: Revolutionizing of IoT based Digital Security Mechanism with Machine Learning Assisted Intrusion Detection

System," 2024 International Conference on Automation and Computation (AUTOCOM), Mar. 2024, doi: 10.1109/autocom60220.2024.10486179.

[7] L. Kangethe, H. Wimmer, and C. Rebman, "Network Intrusion Detection system with Machine learning Intrusion Detection System with Machine Learning As a Service," Journal of Information Systems Applied Research, vol. 17, no. 3, pp. 4–15, 2024, doi: 10.62273/ewql5023..

[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," in *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, Feb. 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," 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, Apr. 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] S. Ifzarne, H. Tabbaa, I. Hafidi, and N. Lamghari, "Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks," Journal of Physics: Conference Series, vol. 1743, no. 1, p. 012021, Jan. 2021, doi: 10.1088/1742-6596/1743/1/012021.

[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. Muthu Kumar, and M. Madhu 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] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/access.2019.2895334.

[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] P. R. Chandre, P. N. Mahalle, and G. R. Shinde. "Intrusion prevention framework for WSN using Deep CNN." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 6 (2021): 3567-3572.

[24] D. Hemanand, G. 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)", Int J Intell Syst Appl Eng, vol. 10, no. 3, pp. 285–293, Oct. 2022.

[25] S. Salmi and L. Oughdir, "CNN-LSTM Based Approach for Dos Attacks Detection in Wireless Sensor Networks," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130497.

[26] 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, Jun. 2020, doi: 10.3390/sym12061046.

[27] 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, 2022, doi: 10.11591/ijece.v12i1.pp914-923.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," in *Machine Intelligence and Pattern Recognition*, 1994. doi: 10.1016/B978-0-444-81892-8.50040-7.
[33] 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.
[34] X. Qi, S. Silvestrov, and T. Nazir, "Data classification with support vector machine and generalized support vector machine," AIP Conference Proceedings, 2017, doi: 10.1063/1.4972718.
[35] 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.
[36] 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.
[37] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, "WSN-DS: A Dataset for Intrusion Detection Systems in Wireless Sensor Networks," *J Sens*, vol. 2016, 2016, doi: 10.1155/2016/4731953.
[38] 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.
[39] H M. Fadhil, N. Q. Makhool, M. M. Hummady, and Z. O. Dawood. "Machine Learning-based Information Security Model for Botnet Detection." Journal of Cybersecurity and Information Management (JCIM) Vol 9, no. 01 (2022): 68-79.
[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, Nov. 2017.
[41] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *Journal of Information Security and Applications*, vol. 44, pp. 80–88, Feb. 2019, doi: 10.1016/j.jisa.2018.11.007.