Synergy: Cross-Disciplinary Journal of Digital Investigation
(ISSN 2995-4827) VOLUME 03 ISSUE 04, 2025

FAULT DIAGNOSIS IN WIRELESS SENSOR NETWORKS USING IMPROVED FIREFLY ALGORITHM AND SUPPORT VECTOR MACHINE

Saif Mohammed hamzah al_asadi*, Zainab abbas alsultani
Department of Computer Technical Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq

Abstract:

Wireless Sensor Network (WSN) consists of a large number of sensor nodes deployed in target areas for specific applications. Reliable transmission of data from the cluster head to the base station for processing is one of the most important challenges to ensure the proper performance of applications in wireless sensor networks. The distinction between good and bad data should be properly realized. Classification of applications in fault detection in wireless sensor network is one of the significant solutions in this field. A new model for detecting the accuracy of collected data based on improved support vector machine (SVM) classification for clustering data into cluster heads in a hierarchical wireless sensor network is proposed in this research. Optimum parameter setting in support vector machine to achieve accurate classification is done through improved firefly algorithm. Environmental information such as temperature, humidity, etc. is transmitted from the sensor nodes to the cluster heads so that the error data is categorized and transmitted to the base station. The simulation results of the proposed method show that the proposed approach provides an effective way to send correct data for WSN applications. This system achieves more than 99% accuracy throughout the data learning process. Data test results proved that the performance of the proposed method is more accurate compared to the base paper method (with accuracy equal to 0.9864).

Keywords: wireless sensor network, firefly algorithm, support vector machine, incomplete data recognition, classification.

1. Introduction

In recent years, the advancement of technology in several research fields such as wireless communication has led researchers to the field of wireless sensor networks. Wireless sensor

network is an innovative technology playing an important role in data processing and is a combination of wireless communication, detection functions and embedded technology. This emerging technology is expanding and has increasingly permeated all aspects of pervasive environmental monitoring and processing. The main feature of these systems is the possibility of deploying them in remote and dangerous places, flexible organization facilities and ease of data access [1].

A wireless sensor network (WSN) consists of a large number of small sensor nodes that are dispersed in a certain geographical area. Although this growing technology combines sensor capabilities, computing and wireless networks, it has low power consumption. After the sensor nodes are deployed in a certain way to create a robust network, they submit their reports to a sink. WSNs have limited power supplies as well as batteries. Basically, sensor nodes cannot be recharged. When you deploy sensor nodes in a dangerous environment, the nodes are vulnerable to attacks and attackers can control the deployment area. In multi-hop networks, processing is done at the intermediate node. Data received from child nodes are aggregated by executing aggregation functions such as SUM, MAX, MIN, etc. The results are then sent to the top level node or to the sink. Data aggregation is an important primary tool for processing, combining and summarizing data packets before transmitting them to higher nodes [2].

Energy conservation is one of the vital factors in this network. Sensors consume a lot of energy when the transmitter sends data. Therefore, packet management in this network is very necessary. This process is done by integrating the data by the intermediate sensors through the network as well as compressing the data. The effects of this process on the energy efficiency of the sensors, increasing the network lifetime and bandwidth are very impressive. In this context, data aggregation is known as an effective technique for combining data. Data collection or aggregation is the process of collecting data from multiple sensors. Another factor affecting data aggregation is the process of data delivery using an efficient method with minimum delay. Therefore, in order to increase the lifetime of the sensor network, various algorithms are introduced to aggregate data according to the conditions [3].

Collections of sensor nodes that are sparsely deployed in inaccessible areas and form data broadcast networks develop a wireless sensor network. The important roles of these nodes are to observe the process, collect data and transmit this data to the base station for final transfer. The idea of sensor networks is based on the collaborative effort between large sets of sensors [4].

A wireless sensor network is an infrastructure-free network that is monitored by a central entity called a base station (BS). These networks are application-specific and the wireless sensor network design depends on this application. Energy limitation is one of the important problems in any wireless network. Energy efficiency is a critical challenge in the design consideration of most wireless sensor network applications. Although the advances in battery technologies lead to the improvement of network lifetime, the development of complex intra-network algorithms is a necessity [5].

If the data is collected and aggregated using aggregation functions like, etc., a large part of the energy (instead of wasting it to send data to the sink node) will be stored individually in the node. Data aggregation in wireless sensor networks includes the process of collecting and combining useful information in the desired area. The effectiveness of communication between nodes depends on the data aggregation technique used. Data aggregation can be considered as a fundamental processing technique to reduce energy consumption and conserve limited resources. Effective data aggregation technique can improve network lifetime and energy efficiency [6].

The technique of developing information identification model for data aggregation in wireless sensor networks based on support vector machine classification optimization and firefly algorithm is introduced in this research. The optimal parameters of the support vector machine and the kernel

function of this machine are the same as the components that evaluate the accuracy of information identification in wireless sensor network applications. The decision-making function in the classifier technique is deployed in the nodes of the hierarchical wireless sensor network to classify the sensor data of mobile nodes and identify their defects for the next steps of processing, i.e. data aggregation. That is, the sensed data from the environmental information are aggregated before being transmitted to the base station through cluster heads. Identification of correct information and detection of data defects are done in the heads of clusters to create the identification model, which in turn is optimized by the improved firefly algorithm.

2. Related Works

Data aggregation based on machine learning for sensor networks is introduced in reference [7]. The network here is built based on the clustering method. Topology contains sensors that are grouped in the form of cluster heads and members. The data and its transmission to the cluster heads are sensed by each member. The American College of Nurse-Midwives or ACNM methodology is then applied to filter the noise and initial errors caused by each sensor by the cluster head. The machine learning based neural network is then activated by the cluster head to train and test the data with learning techniques. The measured data is aggregated by the cluster head before passing to the destination. Packet redundancy and network load are reduced in this packet delivery method in sensor networks. Network nodes accurately manage the entire network resource and minimize network delays and overheads. This reduction increases the life of the network compared to the normal life period. Thus, the network substantially reduces data transmissions and controls the resulting observed latency.

The technique of energy-efficient deep learning-based network slicing with data aggregation (EENS-DA) was developed in [8]. Basically, this method focuses on the clear and efficient allocation of resources needed for a specific application. Also, the EENS-DA model uses convolutional long short-term memory (Conv-LSTM) network cutting techniques and tree-based data aggregation. Criteria such as data cutting effectiveness, accuracy and confidentiality are improved in the EENS-DA technique. The innovation of working with the inclusion of DA in the concept of network slicing in wireless sensor network has been proven.

The hypothesis of rough set is integrated with improved convolutional neural network in reference [9] and a new information aggregation algorithm is proposed for wireless sensor network. First, the feature extraction model is designed in the proposed algorithm, which is trained in the sink node and the rough set hypothesis is implemented in that node to effectively simplify the information and reduce the labeled dimensions. When the data attributes are extracted through the deep network cluster nodes, they are sent by the cluster heads to the sink node to reduce the amount of data transmission and increase the network lifetime. Qualitative and quantitative simulation results have been compared with the results of existing data regression algorithms. The results prove the obvious reduction of energy consumption of the proposed convolutional neural network model and the improvement of data aggregation accuracy.

A new data aggregation model based on intelligent research order techniques is proposed in [10]. First, the research order is formulated and the frames are ranked based on the multi-objective function. A recently developed multi-objective function includes measures of latency, efficiency, and data freshness. A solution similar to the research order is first trained in the neural network using the Fitness-Mated Lion (FM-LA) algorithm. The optimality resulting from the neural network search order is used to generate the second-level solution and is then applied to FM-LA to optimize the next search order. Therefore, the neural network-based two-stage optimization process for sorting research is compared with conventional methods in terms of performance metrics such as latency, efficiency, and data freshness. Therefore, the comparative analysis proves the performance improvement of the proposed model.

A new method for solving security and energy issues in wireless sensor networks named secure lightweight cryptographic data-aggregation algorithm (SLC-DAA) has been introduced in [11]. The proposed SLC-DAA uses cytographic primitives such as hash functions and XOR operations. This method is used to provide promising solutions in data aggregation based on clustering using energy consumption model and higher security. The proposed model is compared with existing simple approaches and provides better computational efficiency and protection of confidential information.

A two-vector data prediction model based on normalized quantile regression (NQR) for data aggregation to minimize data transmission and improve network lifetime through future data estimates based on previously measured interpretations is proposed in [12]. Two-vector data prediction model for data aggregation considering energy efficiency is implemented in this work. The purpose of this method is to organize the predicted and evaluated datasets through sensor nodes and cluster heads to avoid imposing the total error caused by constant predictions. The interpretation of the sensor in the next time slot in the sensor network and the cluster head is predicted based on the data stored in the vectors in the NQR algorithm. When the next data is measured, each sensor network in the cluster compares its predicted data with the actual measured data. If the prediction error is less than a predetermined threshold, the true sensor interpretation for the cluster head will not be used by the sensor network. When the cluster head does not receive a value from the sensor network, it uses the same data prediction algorithm and predicts the next data in a series. As a result, redundant communications do not occur and the energy of the sensor network is conserved. This cycle is not a data transfer cycle. When the prediction error exceeds the preset threshold, the sensor network should send the actual measured data to the cluster head.

A new classification-based pattern for clustering data into clusters using support vector machine classifier in wireless sensor network has been introduced by researchers in [13]. Also, the researchers proposed an advanced version of the flower pollination algorithm (IFPA) to optimize the parameters for the support vector machine. Also, a classification decision function in the wireless sensor network application should be deployed to identify the correct information for conventional data aggregation for the next process. Environmental data collection such as temperature, humidity and other items are categorized into error or normal data classes so that the data is aggregated and sent to the base station. The correct data forwarding pattern in wireless sensor network applications in the proposed method is compared with a series of existing methods. The design of the proposed model system includes several major parts, such as data collection and preprocessing, normalization of data features, and training and testing of data sets. The effectiveness of the effective method of correct data forwarding in wireless sensor network applications has been proven compared to some existing methods.

Researchers in [14] introduced a sensor measurement error detection algorithm based on Pearson correlation coefficient and support vector machine algorithm. Environmental phenomena are spatially and temporally correlated, but errors are not correlated to some extent, so Pearson's correlation coefficient is used to measure correlation. Then, SVM was used to classify and distinguish defective reads from normal reads. After classification, defective reads are discarded. Here, each sensor node periodically collects environmental features and sends them to its associated cluster heads. Each cluster head analyzes the collected data using a classification algorithm and detects errors. Subsequently, NS-2.35 and MATLAB simulators have been used to evaluate the proposed method. The error detection algorithm was evaluated using performance criteria, i.e. accuracy, precision, sensitivity, recall, F1 score, geometric mean, receiver operating characteristic (ROC) and area under the curve (AUC). The performance evaluation process shows that the proposed method has good and stable performance for high percentage errors.

Researchers in [15] have proposed a technique based on supervised machine learning to closely examine the behavior of sensors through their data and detect faults. Most of the errors that commonly occur in WSNs are hardware errors, drift, spike, irregular, data loss, and random errors.

A reliable dataset has been published online by researchers at the University of North Carolina. This data set consists of temperature and humidity sensor measurements in the form of a multi-hop scenario, and the mentioned errors were simulated in non-defective (normal) data. Also, fault events were generated to replicate real WSN scenarios. A lightweight learning-based technique called highly random trees or redundant trees is proposed for timely fault detection and diagnosis.

Researchers in [16] devised the ReliefF algorithm to optimize Canonical Correlation Analysis paired with Improved Support Vector Machine (RCCA-ISVM) for accurate nonlinear circuit diagnostics and fault detection by learning circuit patterns and abnormalities. Initially, input data is gathered from the Sensor Fault Detection database. The Wavelet Packet Transform (WPT) is used to extract fault characteristics from time-domain, statistical, and frequency-domain information. The Canonical connection Analysis (CCA) algorithm is used to improve the connection between the features based on their weights, making the selected characteristics more relevant for defect identification. The fused features are then dimensionally reduced using Principal Component Analysis (PCA). Finally, an ISVM classifier is used to do fault diagnostics based on the reduction.

Researchers in [17] devised a fault diagnosis technique for the entire life cycle that uses wavelet thresholding denoising, genetic algorithm-variational mode decomposition (GA-VMD), and improved grey wolf optimizer-least squares support vector machines (IGWO-LSSVM). The nonlinear convergence factor is used to improve GWO's global search capabilities, while IGWO optimizes LSSVM to improve fault identification accuracy. This technology, which is based on a high-speed rolling bearing test rig, is used to diagnose faults in high-speed rolling bearings throughout their lifetime. The kurtosis and approximation entropy indexes serve as the foundation for stage division in the degradation process. The defect diagnosis of rolling bearings throughout their life cycle is done by five steps: data collecting, stage division, data preprocessing, fault feature extraction, and fault feature identification.

Researchers in [18] conducted a recursive analysis on a vibration signal, and the recursive characteristics were used as a nonlinear recursive feature vector, which included recursive rate (RR), deterministic rate (DET), recursive entropy (RE), and diagonal average length (DAL). Then, a comprehensive multi-domain feature vector is created by merging three time-domain features: root mean square, variance, and peak to peak. Finally, the whale optimization algorithm (WOA) is used to optimize the penalty factor C and the kernel function parameter g, resulting in the best WOA-SVM model.

The proposed detection scheme based on redundant trees has the ability to be robust against signal noises and greatly reduce bias and variance error. The efficiency of the proposed scheme has been compared with advanced machine learning algorithms such as support vector machine, random forest, neural network and decision tree. Performance evaluation proved the effectiveness of the proposed scheme in terms of accuracy, precision and F1 score. In addition, the proposed scheme has less training time compared to the state-of-the-art approaches.

3. The proposed method

In this thesis, we propose a new data aggregation pattern based on clustering in wireless sensor network clusters using improved support vector machine classifier. We also propose an improved version of the firefly algorithm to optimize the parameters for the support vector machine. A classification decision function (due to the need for accurate data) in several successful WSN applications must be deployed to correctly identify information so that routine data can be aggregated for further processing. Figure 1 shows the flowchart of the proposed method.

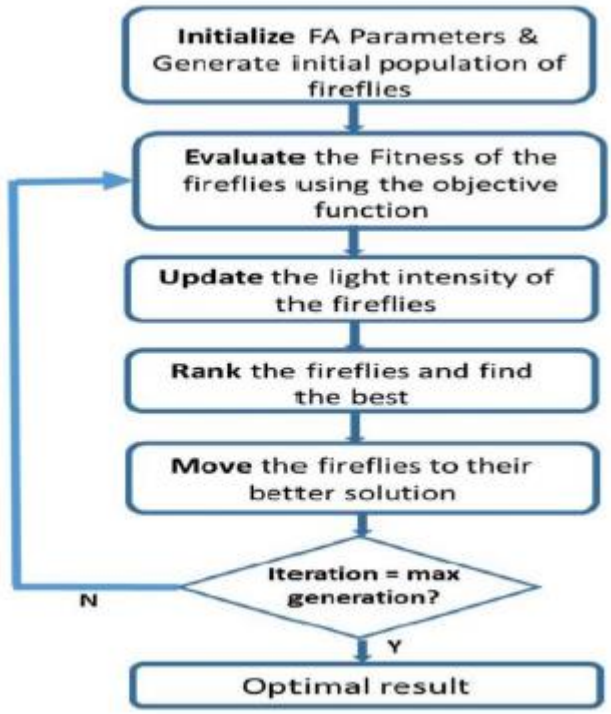

Figure 1- Flowchart of the proposed method

3.1. Support vector machine (SVM)

Achieving a model in order to maximize the data training performance is the main goal in the pattern classification process. Convolution training methods classify each input-output pair into a correct category in specific models. Therefore, if the classifier is highly adapted to the training data, the model will work by memorizing the training data. This challenge reduces the generalization ability of the classifier [19]. The main motivation of SVM is to separate several classes in the training set in such a way that the margin between them is maximized. On the other hand, the ability to maximize model generalization is provided through SVM. The purpose of the structural risk minimization (SRM) rule lies in the fact that it allows minimizing the bounds on the generalization error of the model (instead of minimizing the root mean square error on the training dataset). This method is often used in experimental risk minimization techniques as a working philosophy. The discriminant function of the two-class support vector machine works as follows:

$$f(x) = \sum_{i}^{n}{}_{=1} y_i \alpha_i k(x, x_i) + b \quad (1)$$

N is the number of training samples, x_i is ith training sample and yi is the correct class for ith training sample. The value of y_i is considered to be +1 for one category and -1 for the other category. The bias value of the function is denoted by b. The values are the classification coefficients, denoted by b during training. The $k(x_1, x_2)$ is kernel function of the SVM. Two examples of the most common kernels used for support vector machines are:

- Multidimensional kernel

$$k(x_1, x_2, p) = (1+x_2.x_2)^P \quad (2)$$

- RBF kernel

$$k(x_1,x_2,\sigma)=e^{\frac{-\|x_1-x_2\|^2}{2\sigma^2}} \quad (3)$$

In the above relations P and σ are adjustable parameters for Multidimensional and RBF kernels. By solving the problem, the following optimization objective is achieved.

$$maximize \qquad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i.j}^{n} \alpha_i \alpha_j y_i y_j k(x_j . y_j) \tag{4}$$

In the above relation, the constant C is one of the parameters of the SVM classifier. The effectiveness of SVM depends on the choice of kernel, kernel parameters, soft margin and C parameter. In this research, the RBF kernel function is used for the support vector machine, because the RBF kernel function can analyze highdimensional data and requires only two parameters. These two parameters are obtained through the advanced firefly algorithm.

3.2. Advanced firefly algorithm

The firefly algorithm is one of the heuristic algorithms inspired by the behavior of this insect in nature. Fireflies are one of the most interesting and special creatures in nature. These radiant insects are from the cockroach family and live mainly in tropical and temperate regions. The population of this animal species includes about 1900 distinct species. These insects can emit a lot of light because of the special photo genetic organ located on the surface of the body behind the translucent skin [20]. The firefly algorithm is powerful in the local search process, but because this insect does not have a good global search, it gets stuck in local optima. The parameters of the firefly algorithm may not change in time during successive iterations. Two parameters of the algorithm are "attractiveness coefficient" and "randomization coefficient". The values of these indicators are effective in determining the speed of convergence and the behavior of the firefly algorithm.

Automatic learning machines are adaptive decision-making devices that operate in unknown random environments and progressively improve their performance through a learning process. This type of machine has been successfully used in a variety of applications such as call authorization control in cellular networks, capacity allocation problems, back-propagation parameter matching, and determining the number of hidden units for layered neural networks. Setting firefly parameters in this work is done with automatic machine learning tools with the aim of creating firefly adaptive parameters. A machine learning machine is used to determine the attractiveness coefficient and another one is used to determine the randomization coefficient.

Automatic machine learning is an adaptive decision-making tool that works in random or unknown environments. A machine learning machine has a limited set of actions. Each action has a probability (unknown to the automaton) of getting a reward from the environment. Learning in the context of choosing the optimal action (that is, the action with the highest reward probability) is realized through repeated interaction in the system. If the learning algorithm is chosen appropriately, then a process of iterative interactions in the environment can be performed to select the optimal action.

The variable structure of the automatic learning machine can be shown through the quadruple {α, β, p, T}, where $\alpha=\{\alpha_1, \alpha_2, ..., \alpha_r\}$ is a set of actions of the automatic machine. $\beta=\{\beta_1, \beta_2,\ldots, \beta_m\}$ is equivalent to the set of machine inputs. Also, p={p1, ..., pr} is the probability vector for choosing each action, and p(n +1) = T[α(n),β(n), p(n)] represents the learning algorithm. If β = {0,1}, then the environment is called P-model. If β belongs to a finite set with more than two values between 0 and 1, the environment is called Q-model.

If β is a continuous random variable in the domain [0,1], the environment is called S-model. Suppose VSLA works in the S-Model environment. At the time when action i is executed, the general linear optimization models of action probabilities are represented by the following equation:

$$P_i(n+1) = P_j(n) + a \cdot (1 - p_i(n)) \cdot (b \cdot \beta(n) \cdot p_i(n) \quad (5)$$

$$P_j(n+1) = P_j(n) + a \cdot (1 - p_i(n)) \cdot P_j(n) + (b \cdot \beta(n)) \left[\frac{1}{r-1} - p_i(n)\right] \quad \forall j \quad j \neq i$$

where a and b are equivalent to reward and punishment parameters. When a=b, the automatic machine is called S-LR-P. If b=0 and 0<b<<a, the automaton is called S-LR-I and S-LRϵP.

4.1. Dataset

The data set considered in this study to evaluate the proposed method was collected from 54 Mica2Dot sensors [21] at the Intel Berkeley Research Laboratory between February 28 and April 5, 2004. The sensors are connected to weather panels and information about light, temperature, humidity has been collected with a time stamp. The data is pre-processed with the query processing system. The light is measured by the lux unit. A value of 1 lux corresponds to moonlight, 400 lux corresponds to a bright office, and 100,000 lux corresponds to full sunlight. The measure of humidity is relative humidity corrected by temperature and its value varies between 0 and 100. Temperature is also measured in degrees Celsius.

4.2. Initial setting of parameters

In order to review and evaluate the proposed method, we adjusted our parameters according to the parameters of other papers [13]. The population size of both algorithms is set equal to 50. The switch probability is equal to 0.6. The parameter $R1$ is equal to 10, because the communication must be done every ten iterations. Also, the mutation constant $F0$ is equal to 0.5, the cross-over constant CR is equal to 0.5, the number of the groups Gn is set to 4. Due to the randomness of the meta-heuristic algorithm, the number of runs is set to 200. Table 1 shows the parameters of the proposed method.

Table 1- Initial values for parameters in the proposed method

parameters	values
Training data percentage	80%
Test data percentage	30%
Population size	50
The possibility of a switch	0.6
$R1$	10
Some of the group Gn	4
Number of run	1000
The mutation constant $F0$	0.5
crossover constant CR	0.5

4.3. Experimental results and discussion

In different scenarios in the wireless sensor network, some false nodes in the network to reduce or increase the temperature or humidity unrealistically are simulated. The measured data in wireless sensor network nodes manage traffic congestion conditions. The network is simulated under different conditions and settings with and without errors with variable number of false nodes in the system. Various combinations of false nodes are considered with a variety of one to five nodes. There is an N-node topology. Information packets received by cluster heads are aggregated and sent

to the base station. The data is randomly selected from the member nodes and sent to the cluster heads. Then the cluster heads pack and compress the data and transmit it to the base station node. Under each scenario, there are about 400 data packets randomly generated by sensor nodes.

4.3.1. Environment settings for data collection

It is assumed that the topology of the wireless network has N nodes that are randomly deployed in an area of M×M. The parameter M is equivalent to the area of the establishment land, which can be 200, 300 and 400 meters. N is equivalent to the number of nodes, which varies between 100, 200 or 300. The network has a base station equipped with an unlimited power supply. Aggregated data from cluster heads is received by the base station. The simulation of the proposed method is done in MATLAB software. All runs were completed in the MATLAB 2020b application on a Windows 10 64-bit operating system on an Asus laptop with an Intel Core i7-8665U with 12 GB of RAM. The operational characteristics of the wireless network are set as well as the packet transmission timing periods. The proposed method has been compared with other techniques, i.e. basic paper. The set values for the test parameters are listed in Table 5-1. Collected data includes node ID, packet ID, cluster heads, sensor data (i.e. temperature, humidity, light, network status, gas, etc.) estimated noise and radio signal strength (in decibel units). The sensor data type is also related to the node sensor type in the intended application. The data is divided into two parts: one part for training and another part for testing, with the size of 70% and 30% of the original data set, respectively.

4.3.2. Evaluation criteria

This section discusses the criteria considered to evaluate the classifiers and examines how they perform in predicting class labels. These measures include accuracy, F-score and average g. Before starting to discuss the different criteria, it is necessary to understand some terms. For example, positive examples are examples belonging to a class that are of interest to the user. Subsequently, the negative sample refers to the rest of the samples. Below we present the formulas related to the criteria. Also, the numbers of positive and negative samples are indicated by P and N, respectively.

TP means positive samples that are labeled correctly by the classifier. TN means negative samples that are labeled correctly by the classifier. FN means negative samples whose label is wrongly identified by the classifier. FP means positive samples whose labels are falsely identified by the classifier.

We need criteria to evaluate the quality of the classifier that can distinguish positive and negative samples well. For this, we can use the criteria of resolution and sensitivity. In other words, sensitivity is defined as True positives rate and resolution is also known as True negatives rate.

The sensitivity criterion is defined as follows:

$$\text{sensitivity} = \frac{TP}{TP+FN} \tag{6}$$

The specificity criterion is defined as follows:

$$\text{specificity} = \frac{TN}{TN+FP} \tag{7}$$

Criteria such as precision and mean g are widely used in classification processes. The accuracy criterion means the percentage of samples that have a positive label and in fact their class is positive:

$$\text{precision} = \frac{TP}{TP+FP} \tag{8}$$

The g-mean is also defined as follows:

$$\text{gmean} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{9}$$

The F score criterion is based on the combination of accuracy and sensitivity criteria as below. It is defined as:

$$\text{F-score}=\frac{2\times\text{precision}\times\text{sensitivity}}{\text{precision}+\text{sensitivity}} \tag{10}$$

In addition to expressed criteria terms such as N, P, TP, TN, FP, and FN are summarized in the form of a matrix called the confusion matrix, which is shown in Table 2.

Table 2. Confusion matrix

		Predicted class: yes	Predicted class: no	total
Real class	yes	TP	FN	P
	no	FP	TN	N
	total	P'	N'	P+N

Although the above confusion matrix is designed for the sum of data with two classes, it can be easily extended to data with more than two class labels. Confusion matrix is a useful tool that determines the performance of a classifier for distinguishing samples from different classes. In this matrix, the TP and TN values show the correct performance of the classifier and the FP and FN values show its incorrect performance. If the number of available classes is m, the confusion matrix is a table with a size of at least m × m. The entry $CMi‹j$ in this matrix represents the

number of samples whose real class label is i and the classifier labeled them with class j. An ideal classifier places the majority of samples on the main diameter of the confusion matrix. It is also better that the rest of the elements except the main diameter of the matrix have a value of zero or close to zero. The table representing the confusion matrix uses an additional row or column to display the sum of the values in the matrix. According to the experimental data set and the formulated model, TPR refers to the fraction of positive samples that are correctly labeled by the model, and FPR refers to the fraction of negative samples that are falsely identified as positive by the model. These criteria are actually compiled in the following way:

$$TPrate = \frac{TP}{TP+FN} \tag{11}$$

To prove the classification accuracy of the proposed method, we use two other criteria, namely identification accuracy (IA) and false positive rate (FPR), which are formulated as follows:

$$\text{IA}= \frac{\text{the number of false identified data}}{\text{the total number of false data available}} \tag{12}$$

In this relation, IA is equivalent to detection accuracy, i.e. the ratio of the number of detected false data to the total number of false data available.

$$FPR = \frac{\text{the number of non-false identified data}}{\text{the total number of false data available}} \quad (13)$$

In this relation, FPR is equivalent to the false positive rate, i.e. the ratio of almost non-false data detected to the total number of false data available.

4.3.3. Optimization results

Comparison results of proposed method compared to existing methods (basic paper, [13], [14], [15], [16] and [17] are shown in the Table 3. The comparison of the proposed method (IFPA for optimal parameters of support vector machine and kernel function) based on 5 criteria is shown in Figure 2. According to the results, the best values are provided by the proposed method. The proposed approach helps to significantly improve the FPR compared to other methods. The reason is that the classification is improved by applying IFPA for optimization, SVM parameters and kernel functions. In the tests part, the system has been tested through data collection by IFPA-SVM as an improved classifier, and the test results have been compared with existing methods. The comparison results show that the proposed method is quite effective for sending correct data for WSN applications. This system has achieved an accuracy of more than 99% throughout the data learning process. In data testing, the efficiency of the proposed method is more accurate compared to other methods.

Our improved Firefly algorithm and support vector machine (SVM) method for defect detection in wireless sensor networks outperformed previous methods in a comparative examination. For example, the method provided in Ref [13] with an IA index of 0.9864 and F1 of 0.9856 performs reasonably well, but it is significantly inferior to the proposed method, which has IA and F1 of 0.9989 and 0.9988, respectively. For example, the method provided in Ref [13] with an IA index of 0.9864 and F1 of 0.9856 performs reasonably well, but it is significantly inferior to the proposed method, which has IA and F1 of 0.9989 and 0.9988, respectively. Furthermore, the true positive rate (TPate) published in Ref [13] is 0.9880, but our technique achieves 0.9976, demonstrating a higher ability to detect real defect samples.

Compared to Ref [14], which has a lower IA of 0.8540 and an F1 of 0.7395, demonstrating an imbalance between Precision and Recall, our technique is clearly superior. In Ref [15], despite a Precision of 1, F1 of 0.9333 and Gmean of 0.9354 suggest that the model was unable to reach a perfect balance in recognition, whereas our proposed technique obtained a very favorable balance with F1 of 0.9988 and Gmean of 0.9988.

In terms of overall accuracy, our method (0.9989) exceeds all others; for example, Ref [16] has an accuracy of 0.9930 while Ref [18] has an accuracy of 0.9900. This result highlights the efficacy of integrating the Firefly optimization technique with SVM for extracting effective features and changing parameters properly.

Overall, the findings reveal that the proposed technique not only performs better in detecting faults more precisely, but also significantly beats the reference methods in all important performance metrics such as accuracy, IA, F1, and Gmean. This demonstrates the suggested fault detection system's strong balance, generalizability, and capacity to handle imbalanced data.

Table 3. The results of comparing the proposed method with other methods

Method	Accuracy	IA	TPate	FPR	Precision	F1	Gmean
Ref [13]	-	0.9864	0.9880	0.9850	0.9833	0.9856	0.9865
Ref [14]	-	0.8540	-	-	0.8750	0.7395	-
Ref [15]	-	0.9412	-	-	1	0.9333	0.9354
Ref [16]	0.9930	-	-	-	-	-	-
Ref [17]	0.9700	-	-	-	-	-	-

Ref [18]	0.9900	-	-	-	-	-	-
Proposed method	0.9989	0.9989	0.9976	1	1	0.9988	0.9988

Figure 2 compares the suggested method's accuracy to three approaches accessible from various sources (Refs [16], [17], and [18]). As can be shown, the proposed method outperforms all other methods, with an accuracy of 0.9989.

The method in Ref [16] has the closest performance to our method, with an accuracy of 0.993, but it is still around 0.006 less accurate than the suggested method, which is a considerable difference in sensitive problems like fault detection in wireless sensor networks. In comparison, Ref [17] performs poorly with an accuracy of 0.970, while Ref [18] comes in second with an accuracy of 0.990.

Overall, the findings reveal that the suggested method not only detects problems more accurately, but also strikes a better balance between true fault identification and false alarm reduction. As a result, on the Accuracy bar chart, the "Proposed Method" column should be significantly taller than the other columns, suggesting that our method outperforms the other methods tested.

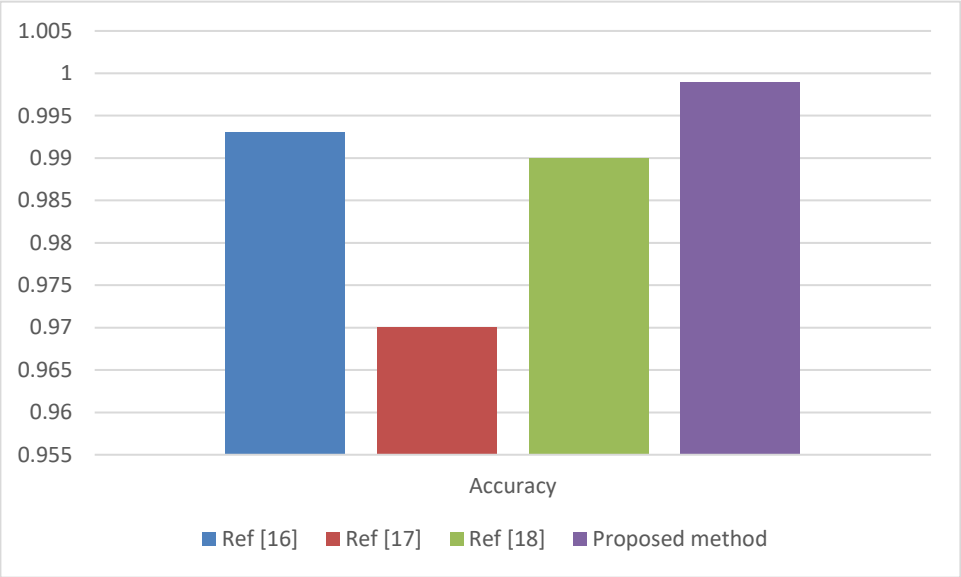

Figure 2. Accuracy Comparison of the Proposed Method with Existing Approaches in Fault Diagnosis for WSN

5. Conclusion

Recent advances in wireless sensor networks have inspired new and diverse applications, including target tracking and ecosystem monitoring. At the same time, a large amount of energy in WSNs is consumed by data communication. However, data aggregation techniques can eliminate redundant data transmitted to the base station and lead to reduced energy consumption. In this thesis, a series of techniques are proposed in the development of information identification model for data aggregation in wireless sensor networks based on optimization of support vector machine classification and improvement of the exploration phase of firefly algorithm. The optimized parameters of the support vector machine and the kernel function of this machine are the same components that confirm the accuracy of information identification in wireless sensor network applications. The decision function in the classifier technique is deployed in the head clusters of the hierarchical wireless sensor network to classify the data measured by the MNs and identify the existing defects for the next steps of processing, i.e. data aggregation. The simulation results of the

proposed method show that the proposed approach is very effective for correct data transmission for WSN applications. The system achieves an accuracy of more than 99% throughout the data learning process. In the data experiments, the performance of detecting incomplete data is better compared to the base paper method.

References

1. Iskandarani MZ. Energy and path loss analysis of wireless sensor networks on a robotic body (WSRobotic). Bulletin of Electrical Engineering and Informatics. 2025 Jun 1;14(3):1794-807.
2. Zolfagharipour L, Kadhim MH, Mandeel TH. Enhance the security of access to IoT-based equipment in fog. In 2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT).
3. Gharaei N, Alabdali AM. Secure and energy-efficient inter-and intra-cluster optimization scheme for smart cities using UAV-assisted wireless sensor networks. Scientific Reports. 2025 Feb 4;15(1):4190.
4. Zolfagharipour L, Kadhim MH. A Technique for Efficiently Controlling Centralized Data Congestion in Vehicular Ad Hoc Networks.
5. Ghosh D. Generation of Key Predistribution Scheme Applying Quasi-Symmetric Designs and Bent Functions in the Wireless Sensor Network. Next-Generation Systems and Secure Computing. 2025 Mar 31:159-85.
6. Karunkuzhali D, Pradeep S, Sungheetha A, Basha TG. Data-Aggregation-Aware Energy-Efficient in Wireless Sensor Networks Using Multi-Stream General Adversarial Network. Transactions on Emerging Telecommunications Technologies. 2025 Feb;36(2):e70017.
7. Sudha, C., D. Suresh, and A. Nagesh. "Accurate Data Aggregation Created by Neural Network and Data Classification Processed Through Machine Learning in Wireless Sensor Networks." (2021).
8. Sheena, B. Gracelin, and N. Snehalatha. "An Energy Efficient Network Slicing with Data Aggregation Technique for Wireless Sensor Networks." In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 13-18. IEEE, 2021.
9. Cao, Junqin, Xueying Zhang, Chunmei Zhang, and Jiapeng Feng. "Improved convolutional neural network combined with rough set theory for data aggregation algorithm." Journal of Ambient Intelligence and Humanized Computing 11, no. 2 (2020): 647-654.
10. Sarode, Prachi, T. R. Reshmi, and Venkatasubbu Pattabiraman. "Combination of Fitness-Mated Lion Algorithm with Neural Network for Optimal Query Ordering Data Aggregation Model in WSN." Wireless Personal Communications 116, no. 1 (2021): 513-538.
11. Kowsalya, R., and B. Roseline Jeetha. "Cluster based data-aggregation using lightweight cryptographic algortihm for wireless sensor networks." Materials Today: Proceedings (2021).
12. Jain, Khushboo, and Akansha Singh. "A Two Vector Data-Prediction Model for Energy-Efficient Data Aggregation in Wireless Sensor Network." (2021).
13. Dao, Thi-Kien, Trong-The Nguyen, Jeng-Shyang Pan, Yu Qiao, and Quoc-Anh Lai. "Identification failure data for cluster heads aggregation in WSN based on improving classification of SVM." IEEE Access 8 (2020): 61070-61084.
14. Biswas, Priyajit, and Tuhina Samanta. "A method for fault detection in wireless sensor network based on pearson's correlation coefficient and support vector machine classification." Wireless Personal Communications 123, no. 3 (2022): 2649-2664.

15. Saeed, Umer, Sana Ullah Jan, Young-Doo Lee, and Insoo Koo. "Fault diagnosis based on extremely randomized trees in wireless sensor networks." Reliability engineering & system safety 205 (2021): 107284.
16. Patchamatla PS, Ramesh M, Abas HM, Pareek PK, Sundaram NK. Fault Detection Using ReliefF Algorithm with Canonical Correlation Analysis Based on Improved Support Vector Machine. In2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS) 2025 Feb 21 (pp. 1-6). IEEE.
17. Li J, Luo W, Bai M, Song M. Fault diagnosis of high-speed rolling bearing in the whole life cycle based on improved grey wolf optimizer-least squares support vector machines. Digital Signal Processing. 2024 Feb 1;145:104345.
18. Wang B, Li H, Hu X, Wang W. Rolling bearing fault diagnosis based on multi-domain features and whale optimized support vector machine. Journal of Vibration and Control. 2025 Mar;31(5-6):708-20.
19. Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. "A comprehensive survey on support vector machine classification: Applications, challenges and trends." Neurocomputing 408 (2020): 189-215.
20. Farahani, Sh M., A. Amin Abshouri, B. Nasiri, and M. R. Meybodi. "Some hybrid models to improve firefly algorithm performance." International Journal of Artificial Intelligence 8, no. 12 (2012): 97-117.
21. Warriach EU, Tei K. Fault detection in wireless sensor networks: A machine learning approach. In2013 IEEE 16th International Conference on Computational Science and Engineering 2013 Dec 3 (pp. 758-765). IEEE.