An innovative machine learning optimization-based data fusion strategy for distributed wireless sensor networks

Naganna Shankar Sollapure, Poornima Govindaswamy
Department of Electronics and Communication Engineering, BMS College of Engineering,
Visveshwarayya Technological University (VTU), Belagavi, India

Article Info

Article history:

Received Feb 29, 2024
Revised Jun 26, 2024
Accepted Jul 14, 2024

Keywords:

Data fusion
DWSNs
GWO
MCSVM
SMOTE

ABSTRACT

Self-sufficient sensors scattered over different regions of the world comprise distributed wireless sensor networks (DWSNs), which track a range of environmental and physical factors such as pressure, temperature, vibration, sound, motion, and pollution. The use of data fusion becomes essential for combining information from various sensors and system performance. In this study, we suggested the multi-class support vector machine (SDF-MCSVM) with synthetic minority over-sampling techniques (SMOTE) data fusion for wireless sensor network (WSN) performance. The dataset includes 1,334 instances of hourly averaged answers for 12 variables from an AIR quality chemical multisensor device. To create a balanced dataset, the unbalanced data was first pre-processed using the SMOTE. The grey wolf optimization (GWO) approach is then used to reduce features in an effort to improve the efficacy and efficiency of feature selection procedures. This method is applied to classify the fused feature vectors into multiple categories at once to improve classification performance in WSNs and address unbalance datasets. The result shows the proposed method reaches high precision, accuracy, F1-score, recall, and specificity. The computational complexity and processing time were decreased in the study by using the proposed method. This is great potential for accurate and timely data fusion in dispersed WSNs with the successful integration of data fusion technologies.

Corresponding Author:

Naganna Shankar Sollapure
Department of Electronics and Communication Engineering, BMS College of Engineering
Visveshwarayya Technological University (VTU)
Bengaluru-560019, Karnataka, India
Email: naganna.sollapure@gmail.com

1. INTRODUCTION

The field of industry and research indicates a growing interest in wireless sensor networks (WSNs). WSN is essentially a node-based network that works together to perceive and potentially control their surroundings, allowing people or computers to interact with the environment. WSNs open up new applications and potential markets, but they also impose certain limits on design that need the development of new paradigms [1]. Detecting, processing, and communicating while using limited energy spurs a cross-layer design approach that usually necessitates taking medium access control, communication protocols, and distributed signal/data processing into account simultaneously. WSNs are essential to the growth of the internet of things (IoT), and communication speed is a key component of both [2].

WSN is capable of managing intricate situations, collecting a variety of information, and adapting to various different settings. Detecting radioactive radiation, chemical and biological pollutants, and guiding soldiers into dangerous areas are just a few of its numerous militaries uses [3]. In addition to other environmental monitoring and protection tasks, it may be used to gather data in the field, follow animal tracks, evaluate pollution levels, and forecast the occurrence of streams of debris and forest fires [4]. Following crop development and product flow is the foundation of intelligent production in both industry and agriculture. Several sensor nodes are frequently seen in a WSN. Sensors can establish communication connections with other nodes through radio emissions [5]. Power components, radio transmitters, storage, and computing units are all embedded in it. A single WSN node has limited memory, communication bandwidth, processing speed, and other features. Following injection, the sensor node is in charge of self-organizing within the system's multi-hop communication procedure and appropriate network structure [6]. For the purpose of developing protocols, WSNs can adapt new methods and get unconventional approaches due to a variety of circumstances. It is necessary to find a suitable balance between signal and data computation capabilities since a longer network lifespan requires minimal complexity and energy consumption. It causes scientific phenomena to provide their greatest energy.

Although the domain is restricted to straightforward reporting and data-driven domains, WSN has seen a variety of improvements lately in the areas of energy and computationally efficient procedures. One of the most significant and vital uses of WSNs is monitoring. To detect and exchange physical or environmental characteristics together, such networks employ a vast number of electronic sensors dispersed over the area of interest (AOI). Random generation is usually used to decide the initial placements of the sensor nodes instead of predefined or planned locations [7]. The WSN system provides safety in transportation by assessing potential hazards in unattended regions and transmitting data packets to a central base station. Because of its adaptability, it is crucial in a variety of disciplines, including medical, smart homes, surveying, military reconnaissance, agricultural irrigation, industrial monitoring, and smart transportation. These professions depend on it to provide safe and effective decision-making in remote places [8]. Sensory data is crucial for smart environments, and smart technologies, including digital electronics and wireless communication, enable the development of WSNs. These networks collect environmental information like temperature, pressure, humidity, and gas presence. However, resource constraints like limited computing capacity and energy must be balanced for optimal network performance [9]. The data processing, transmission, control, and storage in complex networks have all seen major changes as a result of automation and wireless communication technology advancements. Intelligent networks are very valuable for research and have great application potential. They combine wireless data transfer, intelligent learning, data-awareness, and dynamic optimization. WSNs, cognitive radar, cognitive radio networks, multi-robot networks, smart grids, and cognitive network sensing are a few examples of intelligent networks [10]. In this study, we discussed a distributed wireless data fusion from sensor networks strategy based on machine learning optimization.

Contributions to this paper includes: i) the dataset includes 1,334 instances of hourly averaged answers for 12 variables from an air quality chemical multisensor device; ii) the dataset was pre-processed using SMOTE. This technique transformed imbalanced data into a balanced dataset; iii) the GWO algorithm for feature reduction is an effort to improve the efficacy and efficiency of feature selection procedures; and iv) according to the results, the suggested strategy achieves superior recall, accuracy, precision, and F1-score. The remainder of the paper is arranged as follows: a literature review is presented in section 2. In sections 3 and 4 present the methodology and results. Finally, section 5 presents the conclusion.

2. LITERATURE REVIEW

2.1. Deep learning-based WSN

An unsupervised distributed multitask estimating technique with asynchronous data-driven adaptive cluster learning is suggested to address these issues and produce an estimate that is more accurate. Hua *et al.* [11] proposed the extensively considers and examines the time delay and different sample rates. The mean stability, adaptive cluster learning behaviour, and mean-square convergence of the proposed method are investigated with asynchronous data. To attain the optimum balance of load and energy efficiency at the fusion center of WSNs, Al-Nader *et al.* [12] provided a deep learning-based distributed data mining (DDM) model. Recurrent neural networks with long short-term memory (RNN-LSTM) are part of the DDM model that is provided. RNN-LSTM separates the network into many layers and places them in the sensor nodes. Together with fewer data transmissions, the suggested model lowers overhead at the fusion center. The suggested the RNN-LSTM model is evaluated in a range of experimental conditions, such as signaling intervals and the quantity of nodes in the hidden layer. Zhuang *et al.* [13] presented a detailed investigation of multi-sensor data fusion, which has been used to integrate positioning/navigation systems within the past decade. This article categorises and expounds on various navigation/positioning systems from three perspectives: sources, algorithms and designs, and situations. We address the Kalman filter and its derivatives, graph optimization techniques, and integrated approaches for analytics-based fusion. Dubey *at al.* [14] offered proximal policy optimization-based ant colony optimization (PPO-ACO) algorithm is a new approach for optimal path selection in WSN. It combines PPO with swarm intelligence techniques like ACO and reinforcement learning to address the challenging trade-off between energy efficiency and security. Simulation data shows the algorithm outperforms previous ones and shows a decrease in average residual energy. Gupta [15] developed the convolution neural networks (CNNs) with gated recurrent units (GRU) model, which integrates gates for recurrent units with convolution units for human activity detection. The technique has proven to be more accurate than other models, such as inception time and deep convolution LSTM, and has been successfully verified on the WISDM dataset.

2.2. Data fusion approach for WSNs

Guezouli *et al.* [16] provided the process of data merging for mobile and diverse WSNs. It treats WSN nodes as the neural network neurons of advanced learning engines. To significantly minimize the amount of data on the network provided to the node of sink, the machine's neural network for extremely high learning collects the sensual information gathered by the mobile WSNs with heterogeneity and merges it with the clustering path. Vidya and Sasikumar [17] provided a supervised machine learning (ML) based activity detection system leveraging multi-resolution time-frequency evaluation of received signal strength (RSS) across wearable sensors. To identify the important feature vector, the multi-sensor activity information obtained from the inertia sensors integrated in a smartphone and WSN hubs is separated utilising the discrete wavelet transform (DWT) and empirical mode decomposition (EMD) techniques. Himeur *et al.* [18] offered a thorough analysis of the data fusion techniques now in use to curb excessive consumption and advance sustainability. Along with doing a taxonomy of current data fusion methodologies and other relevant elements, we look at their conceptualizations, benefits, problems, and downsides. Lavanya and Shanmugapriya [19] introduced intelligent data fusion techniques (IDFTs) may minimize unnecessary data by a large margin, lower the amount of data sent, increase bandwidth efficiency, prolong the life of the network, and eventually ease energy and bandwidth constraints. The paper proposes intelligent data fusion using improved whale optimisation algorithms (IWOAs) to enhance data dependability. By offering insufficient data, IWOAs improve the standard of information obtained from sensor sources while reducing the amount of data collected. An energy-efficient routing protocol utilizing an established adaptive algorithm with a crossover manage, an encoding system, and a modified mutation operations to aid in node identification was proposed in the study [20].

In addition, a highly optimized movement method was implemented in the framework that was suggested for data fusion processes. This technique helps determine just the optimal energy networks and removes redundant alternatives that help improve energy efficiency and additionally minimize transfer of data. A multi-hop construction serves as the foundation for a better backward propagation neural network. For mobile heterogeneous WSNs, the study [21] analyzed a data fusion approach using artificial cell swarm optimization is recommended. The model adapts the neural network of the extreme learning machine (ELM) to control inconsistent output while decreasing information transfer. To optimize WSN lifespan, the study [22] used a sensor deployment technique involving a number of approaches, such as entropy-driven data aggregation with gradient distribution (EDAGD). The strategy combines gradient implementation, multi-hop tree-based gathering of data, entropy-driven consolidation tree-based routing, and gradient distribution. The results show that EDAGD is more efficient than standard random deployment methods. The combined approach for internal navigation presented in the paper [23] combined Wi-Fi RTT with pedestrian dead reckoning (PDR). An exception identification tactics, an adaptive filtering framework, a technique for fusion utilizing observation and federated filter (FF), and a real-time averaging strategy with set intervals are all used in the method. Dautov *et al.* [24] suggested a data fusion in IoT, particularly in smart healthcare. To provide fast and accurate results, it suggests a distributed hierarchical data fusion architecture that combines several data sources at each level. When making judgments that are crucial to the objective, this method reduces waiting times. Utilizing complex event processing technology, the method manages real-time data streaming, which is necessary for IoT devices with limited storage and time-sensitive applications.

3. METHOD

In this section we discuss the data preparation, reduction, and classification are part of the data fusion approach for distributed WSNs. Figure 1 shows the flow of the research. Through the setup air quality data are aggregated and data preprocessing can be done using SMOTE. After preprocessing the feature reduction can be done using GWO approach. Further classification can be done using SDF-MCSVM.

3.1. Dataset

An air quality multisensor device implanted with 1,334 occurrences of hourly averaged responses for 12 variables makes up the dataset. Records of data were made from April 2010 to September 2022 (twelve years) indicating the longest accessible records of responses from field-deployed air quality chemical sensor devices. Table 1 depicts the sample of the dataset. The dataset is distributed into two phases training and testing. Training phase 80% and testing phase 20%.

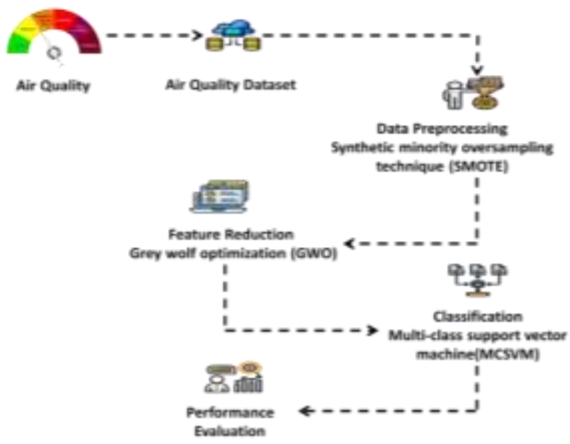


Figure 1. Flow diagram of the study

Table 1. Sample dataset

SAMP _DAY	MONITOR STN	SO2_ pre-processed	NO2_ pre-processed	PM10_ pre-processed	SPM_ pre-processed	PM10_ sub-index	S02_ sub-index	N02_ sub-index	SPM_ sub-index	AQI	AQI-BUCKET
NaT	{‘Chavara’}	80	24	225	565	183.33	100	30	417.85	418	3
NaT	{‘Chavara’}	40	24	230	600	186.67	50	30	442.85	443	3
NaT	{‘Chavara’}	60	32	145	495	130	75	40	367.85	368	4
01/13 /2010	{‘Chavara’}	60	52	285	610	235	75	65	450	450	3
01/15 /2010	{‘Chavara’}	160	30	355	420	306.25	126.67	37.5	314.28	314	4

3.2. Data pre-processing

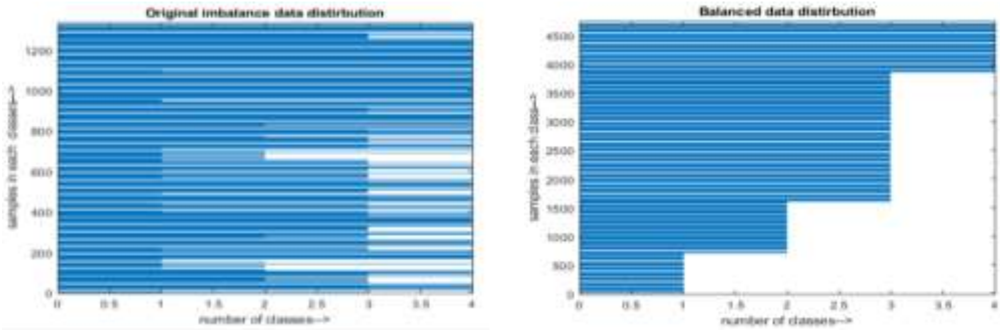
The gathered data was pre-processed in two stages. During the first stage, two variables were neglected, resulting in a dataset comprising a total of nine sensor data variables. In the second stage, the SMOTE was applied to address the imbalance in the dataset, transforming it into a balanced dataset. In step 1, the two variables are eliminated, as shown in Table 2.
Stage 1: In this stage, the data is processed two variables are removed.
Stage 2: SMOTE.

Table 2. Stage 1 database

SO_2_Pre-processed	NO_2_Pre-processed	PM10_Pre -processed	SPM_Pre-processed	PM10_sub index	SO_2_sub index	NO_2 sub index	SPM Sub index	AQI	AQI BUCKET
80	24	225	565	183.33	100	30	417.86	418	3
40	24	230	600	186.67	50	30	442.86	443	3
60	32	145	495	130	75	40	367.86	368	4
60	52	285	610	235	75	65	450	450	3
160	30	355	420	360.25	126.67	37.5	314.29	314	4

A balanced dataset performs better in terms of classification than an unbalanced dataset. As a consequence, a variety of sampling techniques were put forth in an attempt to correct the unbalanced dataset by altering the sample distribution. To enhance the amount of minority class samples, one possible approach is to employ random over-sampling via randomly resampled samples; however, this approach can lead to overfitting. The most popular and efficient oversampling approach is called SMOTE. To create minority samples and balance the unbalanced dataset, it synthesizes additional minority specimens by discovering many neighbors in each minority class sample’s same class. Even while oversampling might reduce the unbalanced ratio, a significant number of synthetic samples can dilute the original minority class samples.

Rebalancing the dataset involves downsampling, which removes out a chunk of most class samples. Random downsampling may cause a loss of some important and distinctive samples. Figure 2 presents the use of SMOTE; Figure 2 (a) shows the imbalanced data distribution and Figure 2(b) represents the balanced data distribution after SMOTE. By interpolating between current minority class samples, SMOTE creates synthetic samples for the minority class. This makes the dataset better suited for training classifiers that could be biased towards the majority class by helping to maintain a balanced distribution of classes. Increase the size of the minority class samples.



(a) (b)

Figure 2. After using SMOTE (a) imbalanced data and (b) balanced data

3.3. Feature reduction using grey wolf optimization

After the pre-processing, the data was reduced by the feature using GWO. The algorithm classifies wolves into four groups based on the hierarchy and cooperation that wolves exhibit when hunting prey. The method is the solution for engineering optimization procedure as a hunting with gray wolves' technique by referring concerns the allocation of food and hunting territories among gray wolves in the wild, using real wolves in place of the main body and a collaborative route search framework based on a separation of responsibility.

Considering that the wolf pack is composed of M people and looks for a meal in a place with dimensions of D the location of the j^ the expression in mathematics which can represent a gray wolf that hunts for food $W\_j = (W\_j\text{^}1, W\_j\text{^}2, \ldots, W\_j\text{^}C\ )$, $j = 1,2, \ldots, M$, when the restriction M is the size of the population of gray wolves. Additionally, it is specified that α is the prey where the best solution can be located, and so forth, β is the second-best solution, δ is the third best solution, with the remaining solution is ω. Each gray wolf in the pack has a defined number as W, and the locations of the wolves' placements α, β, and δ inside the wolf pack that the solution through iteration produced are W_α, W_β, and W_δ to confirm each other's locations. In (1) serves as their calculating formula.

$$\begin{aligned} W_{j,\alpha}^{c}(s+1) &= W_{\alpha}^{c}(s) - B_{j,1}^{c}\left|D_{j,1}^{c} W_{\alpha}^{c}(s) - W_{j}^{c}(s)\right| \\ W_{j,\beta}^{c}(s+1) &= W_{\beta}^{c}(s) - B_{j,1}^{c}\left|D_{j,1}^{c} W_{\beta}^{c}(s) - W_{j}^{c}(s)\right| \\ W_{j,\delta}^{c}(s+1) &= W_{\delta}^{c}(s) - B_{j,1}^{c}\left|D_{j,1}^{c} W_{\delta}^{c}(s) - W_{j}^{c}(s)\right| \end{aligned} \tag{1}$$

In this case, the parameter s is the number of the current iteration the parameters $W_\alpha,\ W_\beta, and\ W_\delta$ are the locations of the wolves in their present range to their prey $B_j^c\left|D_j^c W_\alpha^c(s) - W_j^c(s)\right|$ is the circumferential step width, and the convergence factor computation formulas B_j^c the swing element D_j^c consist of the (2).

$$B_j^c = 2b \times rand - b,\ D_j^c = 2rand_2 \tag{2}$$

The formula for calculating parameter b is (3).

$$b = 2 - \frac{2s}{s_{max}} \tag{3}$$

A computation period for the gray wolf population is shown by the parameter c in this instance and s_{max} signifies the iterative termination timings for the individual gray wolf in the repeated calculation.

$$W_j^c(s+1) = \sum_{i=\alpha,\beta,\delta} \omega_i W_{j,i}^c(s+1)\ \mathrm{C} \tag{4}$$

Here, $\omega_i(i = \alpha, \beta, \delta)$ symbolizes the specific gray wolf α, β, and δ weight coefficient, which can be calculated in the way that follows.

$$\omega_i = \frac{e(W_i(s)}{e(W_b(s))+e\left(W_\beta(s)\right)+e(W_\delta(s))} \tag{5}$$

Now, $e(W_i(s)$ symbolizes the level of fitness of the i^{th} lone wolf pack member in the s^{th} group.

The complexity of time implicitly represents the duration of the algorithm's execution. It is anticipated that the GWO method takes w_1 the duration of execution for setting the parameters and w_2 processing duration to create a consistent distribution. The first step of the GWO algorithm has the following temporal complexity: $e(m)$ is the duration required to ascertain the fitness value.

$$P(w_1 + M\left(mw_2 + e(m)\right) = P(m + e(m)) \tag{6}$$

Taking into the iterative process's execution time upgrading of a single dimension is equivalent to w_3, this indicates that the amount of time required considering the benefits and drawbacks to choose the best solution after several iterations is w_4. At this stage, the temporal algorithm's complexity as (7).

$$P\left(M\left(mw_3 + e(m)\right) + w_4\right) = P\left(m + e(m)\right) \tag{7}$$

Therefore, the GWO procedure's overall temporal complexity is follows: in this study, a dataset comprising 1,334 instances and 12 variables was collected. Feature reduction was performed using the GWO, resulting in the selection of 9 variables for subsequent analysis. Algorithm 1 depicts the pseudocode of GWO.

$$S(m) = P\left(m + e(m)\right) + P\left(m + e(m)\right) = P\left(m + e(m)\right) \tag{8}$$

Algorithm 1. Pseudocode for GWO
```
Establish the population of gray wolves W_j(j = 1,2,……,M
Initialize b, B, and D
Determine each search agent's fitness.
W_α=the best search agent
W_β=the second best search agent
W_δ=the third best search agent
While (s < max number of iterations)
Regarding every search agent
Adjust the current search agent's location by approximate formula
End for
Update b, B, and D
Compute W_α,W_β and W_δ
s = s + 1
End while
Return W_α
```

3.4. Classification using SMOTE data fusion with multi-class support vector machine

The SDF-MCSVM technique is a convenient method for handling classification problems in conditions with multiple classes and unbalanced datasets. It manages sparse data representations and reduces class imbalances, making it useful in real-world scenarios and providing a balanced trade-off between computing efficiency and classification accuracy. Furthermore, one should take into account the computing resources and temporal complexity.

3.4.1. Multi-class support vector machine

MCSVM is a popular machine learning method because of its enhanced functionality. Examine the issue of categorizing information inside a data space W any of the two classes G or: $\neg G$ (not G). Assume that every point of data W possesses a feature vector $\vec{w}$ inside a certain feature area $W \subseteq \Re^m$. We receive l data points $w_1, w_2, \dots, w_l$, referred to as training points, with labels $z_1, z_2, \dots, z_l$, individually. We must forecast whether a fresh data point W is in G or not. MCSVMs are a productive way to address this issue. If the data space is finite, the following are the usual stages in MCSVM: the restricted optimization issue that follows may be used to develop the concept of SVM.

$$\min_{x, a, \xi} \frac{1}{2}||x||_2^2 + D\sum_{j=1}^{m} \xi j \tag{9}$$

Subject to $z_j(x.w_j + a) \geq 1 - \xi j$

$$\xi j \ \geq 0, j = 1, \dots, m \quad (10)$$

- Describe a kernel function $L: W \times W \rightarrow \Re$. there must be symmetry in this function, and the $l \times l$ matrix $[L(w_j, w_i)]_{j,i=1}^{l}$ it needs to be semi-definite and positive.
- Optimize:

$$X(\alpha) = \sum_{j=1}^{l} \alpha_j - \frac{1}{2}\sum_{j,i=1}^{l} z_j z_i \alpha_j \alpha_i L(w_j, w_i) \quad (11)$$

$$\sum_{j=1}^{l} z_j \alpha_j = 0 \quad (12)$$

$$0 \leq \alpha_j \leq D, j \in [1, l] \quad (13)$$

Assume that $\{\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*\}$ is the answer to this optimization issue. We decide $a = a^*$ so that $z_j g_L(w_j) = 1$ for all i with $0 < \alpha_j^* < D$. The training points that are associated with such (j, α_j^*) are referred to as the $support\ vector$. The goal has three terms a regularization term that measures the separation margin inversely, a second term that evaluates the training error, and a hyperparameter C>0 that indicates the trade-off among the two terms.

$$\max_{\alpha} \sum_{j=1}^{k} \alpha_j - \frac{1}{2}\sum_{j,i=1}^{k} \alpha_j \alpha_i z_j z_i\, w_j . w_i$$

Subject to,

$$\sum_{j=1}^{k} \alpha_j z_j = 0, 0 \leq \alpha_j \leq D, \forall\, j \quad (14)$$

the decision function for a fresh input vector x is provided by: when optimum α have been determined.

$$E(w) = sgn(\sum_{j=1}^{k} \alpha_j z_j\, w_j . w + a) \quad (15)$$

The classification determination rule for an information point $w\ \ is\ \ w \in H$ if and only if$sign\ (g_K(w)) = 1$, where,

$$g_L(w) = \sum_{j=1 \rightarrow l, w_j\ \ is\ a\ support\ vector} \alpha_j^* z_j L(w, w_j) + a^* \quad (16)$$

$$l(w_j, w_i) = \exp\ (\frac{-||w_j - w_i||^2}{2\sigma^2} \quad (17)$$

Mercer's theorem states that a feature space is real $\vec{W}$ where the kernel is located K described previously is the internal result of $\vec{W}$ $(that\ is\ L(w, y) =< \vec{w}.\vec{y} > for\ every\ w, z \in W)$. The function $g_L(.)$ denotes the superplane inside $\vec{W}$ that keeps the training locations as far apart as possible in $\vec{W}$ (G points on the plane's positive side and: $\neg G$ points away from the positive). When applied to the test data, it is demonstrable that MCSVM has a limited classification error.

4. RESULT AND DISCUSSION

In this study we used the multi-class SVM combined with SMOTE data augmentation can improve classification accuracy, particularly in unbalanced WSN datasets. SDF-MCSVM ensures accurate and dependable classification across all classes by addressing class imbalance issues that are frequently encountered in WSN applications. The accuracy, precision, recall and F1-score are improved in the suggested approach as compared with existing approaches.

4.1. Experimental result

The suggested approach is implemented in the experiments using MATLAB. MATLAB is a popular modeling and analysis program that is utilized in many different systems and applications. By doing analysis utilizing a vast amount of data sets, in fact aid in extending the ideas behind solutions beyond the desktop.

4.2. Evaluation criteria

The proposed method was compared with three existing methods: data fusion approaches (DFA) [25], context-aware data fusion technique (CDFT) [25], and hybrid learning classifier model (CDFT-HLCM) [25]. The evaluation metrics such as specificity, F1-score, recall, accuracy, and precision. Table 3 indicates the overall results of the recommended and current methods.

$$Accuracy \rightarrow \frac{True\ Postive + True\ Negative}{True\ Postive + False\ Postive + False\ Negative + True\ Negative} \tag{18}$$

$$Precision \rightarrow \frac{True\ Postive}{True\ Postive + False\ Postive} \tag{19}$$

$$F1 - Measure \rightarrow 2 * \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{20}$$

$$Recall \rightarrow \frac{True\ Postive}{True\ Postive + False\ Negative} \tag{21}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Postive} \tag{22}$$

The proposed method was compared with three existing methods: DFA [16], CDFT [16], and CDFT-HLCM [16]. The evaluation metrics are such as F1-score, recall, accuracy, and precision. Table 3 indicates the overall results of the recommended and current methods.

Table 3. Outcomes of the suggested and existing approaches

Methods	Recall (%)	F1-score (%)	Precision (%)	Accuracy (%)
DFA [16]	88.19	85.67	83.29	86.05
CDFT [16]	90.86	89.88	88.92	90.95
CDFT-HLCM [16]	94.67	93.83	93.01	94.4
SDF-MCSVM [proposed]	96.23	96.57	97.58	98.65

4.2.1. Accuracy

The accuracy parameter has a critical function in fine-tuning the dependability and quality of the data fusion using synthetic data procedures for dispersed WSNS. The DFA, CDFT, and CDFT-HLCM only succeed with 86.05%, 90.95%, and 94.40% accuracy, respectively, the suggested technique SDF-MCSVM achieves 98.65% accuracy. Figure 3 depicts the accuracy of the proposed and existing methods.

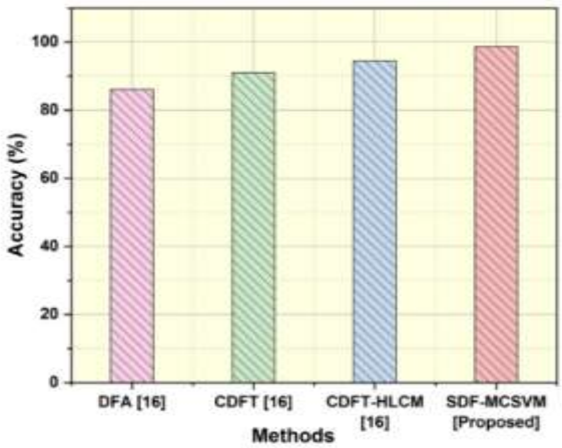


Figure 3. Accuracy comparison with the proposed and existing methods

4.2.2. Precision

The precision parameter reflects contextual sensitivity, which constantly modifies the fusion process to account for changing network dynamics, sensor locations, and environmental changes. The recommended method SDF-MCSVM achieves 97.58% precision, while DFA, CDFT, and CDFT-HLCM manage 83.29%, 88.92%, and 93.01% precision, respectively. Figure 4 displays the precision of the existing and proposed approaches.

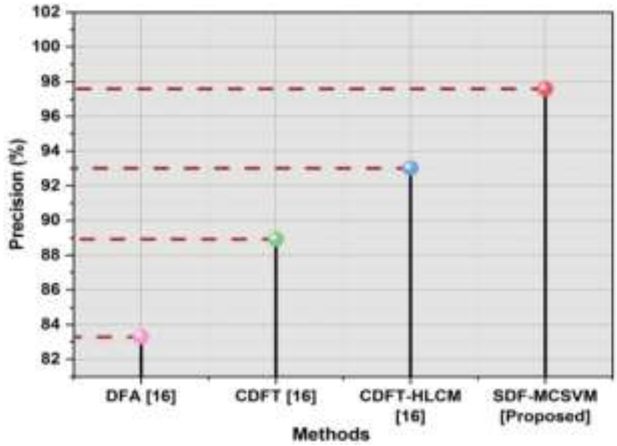


Figure 4. Precision comparison of SDF-MCSVM with existing methods

4.2.3. Recall

The recall parameter adjusts the fusion process dynamically to take into consideration differences in data availability, sensor reliability and environmental dynamics across various time frames and places. This parameter is in line with the concept of temporal and geographical context. The recommended method SDF-MCSVM achieves 96.23% recall, while DFA, CDFT, and CDFT-HLCM manage 88.19%, 90.86%, and 94.67% recall, correspondingly. Figure 5 presents the proposed and existing methods of recall.

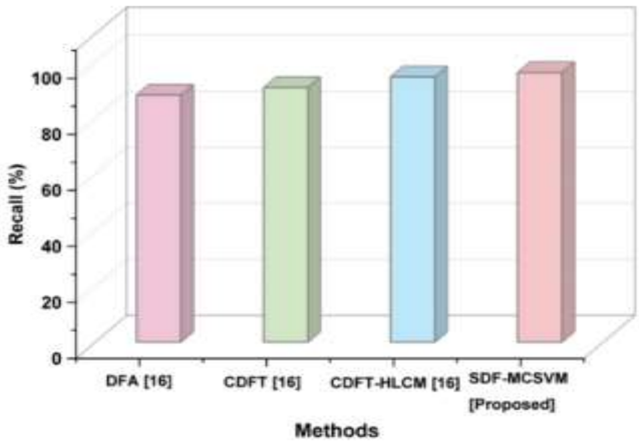


Figure 5. Recall of the proposed and existing methods

4.2.4. F1-score

The F1-score enhances the precision and pertinence of the data by guaranteeing that the fusion process stays flexible and sensitive to changing circumstances. The recommended method SDF-MCSVM achieves 96.57% F1-score, while DFA, CDFT and CDFT-HLCM manage 85.67%, 89.88%, and 93.83% F1-score, respectively. By defining positive and negative outcomes, we can evaluate the effectiveness of the data fusion approach. Given that all circumstances, where the genuine outcome is negative which is taken into consideration, in this context would refer to the fusion algorithm's ability to identify scenarios where the outcome is negative. Figure 6 denotes the F1-score of the proposed and existing methods.

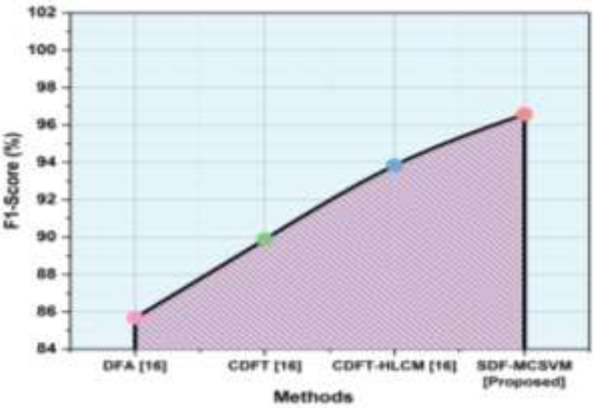


Figure 6. The proposed and existing techniques F1-score

5. CONCLUSION

In this study, the data fusion strategy offers an efficient solution for managing and processing data in distributed WSNs. The air quality dataset obtained 12 variables. The original dataset is pre-processed two stage second stage using the SMOTE approach. The SMOTE technique is utilized to transform imbalanced data into balanced data by synthesizing new minority class instances. By employing GWO, feature reduction intends to increase interpretability, decrease computational complexity, and improve model performance. Classification MCSVM evaluated 9 variables. The result shows the proposed SDF-MCSVM method has better than accuracy (98.65%), recall (96.23%), precision (97.58%), and F1-score (96.57%) existing methods. This proposed method used in the study reduced the computational complexity and processing time. In the future to combine data from several sensors, employ ensemble learning strategies like stacking models, gradient boosting machines, and random forests. Ensemble approaches increase overall performance and reliability by combining the predictions of several base models, particularly in situations where individual sensors may show ambiguity or unpredictability.

REFERENCES

[1] M. Chen *et al.*, "Distributed learning in wireless networks: recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021, doi: 10.1109/JSAC.2021.3118346.

[2] S. Sharma and V. K. Verma, "An integrated exploration on internet of things and wireless sensor networks," *Wireless Personal Communications*, vol. 124, no. 3, pp. 2735–2770, 2022, doi: 10.1007/s11277-022-09487-3.

[3] S. A. Aldalahmeh and D. Ciuonzo, "Distributed detection fusion in clustered sensor networks over multiple access fading channels," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 317–329, 2022, doi: 10.1109/TSIPN.2022.3161827.

[4] X. Xu, J. Tang, and H. Xiang, "Data transmission reliability analysis of wireless sensor networks for social network optimization," *Journal of Sensors*, vol. 2022, pp. 1–12, Jan. 2022, doi: 10.1155/2022/3842722.

[5] J. Logeshwaran and T. Kiruthiga, "Interference-resistant communication framework for sensor nodes in wireless sensor networks," *International Journal of Research In Science & Engineering*, no. 23, pp. 61–75, 2022, doi: 10.55529/ijrise.23.61.75.

[6] M. Hosseinzadeh *et al.*, "A cluster-based trusted routing method using fire hawk optimizer (FHO) in wireless sensor networks (WSNs)," *Scientific Reports*, vol. 13, no. 1, p. 13046, Aug. 2023, doi: 10.1038/s41598-023-40273-8.

[7] A. Boualem, C. De Runz, M. Ayaida, and H. Akdag, "A fuzzy/possibility approach for area coverage in wireless sensor networks," *Soft Computing*, vol. 27, no. 14, pp. 9367–9382, 2023, doi: 10.1007/s00500-023-08406-3.

[8] A. A. Qaffas, "Optimized back propagation neural network using quasi-oppositional learning-based African vulture optimization algorithm for data fusion in wireless sensor networks," *Sensors*, vol. 23, no. 14, p. 6261, Jul. 2023, doi: 10.3390/s23146261.

[9] D. Wohwe Sambo, B. O. Yenke, A. Förster, and P. Dayang, "Optimized clustering algorithms for large wireless sensor networks: a review," *Sensors*, vol. 19, no. 2, p. 322, Jan. 2019, doi: 10.3390/s19020322.

[10] F. Tan, "The algorithms of distributed learning and distributed estimation about intelligent wireless sensor network," *Sensors*, vol. 20, no. 5, p. 1302, Feb. 2020, doi: 10.3390/s20051302.

[11] Y. Hua, H. Gan, F. Wan, X. Qing, and F. Liu, "Distributed estimation with adaptive cluster learning over asynchronous data fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 5, pp. 5262–5274, 2023, doi: 10.1109/TAES.2023.3253085.

[12] I. Al-Nader, A. Lasebae, R. Raheem, and A. Khoshkholghi, "A novel scheduling algorithm for improved performance of multi-objective safety-critical wireless sensor networks using long short-term memory," *Electronics*, vol. 12, no. 23, p. 4766, Nov. 2023, doi: 10.3390/electronics12234766.

[13] Y. Zhuang *et al.*, "Multi-sensor integrated navigation/positioning systems using data fusion: from analytics-based to learning-based approaches," *Information Fusion*, vol. 95, pp. 62–90, 2023, doi: 10.1016/j.inffus.2023.01.025.

[14] G. P. Dubey *et al.*, "Optimal path selection using reinforcement learning based ant colony optimization algorithm in IoT-Based wireless sensor networks with 5G technology," *Computer Communications*, vol. 212, pp. 377–389, 2023, doi: 10.1016/j.comcom.2023.09.015.

[15] S. Gupta, "Deep learning based human activity recognition (HAR) using wearable sensor data," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100046, Nov. 2021, doi: 10.1016/j.jjimei.2021.100046.

[16] L. Guezouli, K. Barka, and A. Djehiche, "UAVs's efficient controlled mobility management for mobile heterogeneous wireless sensor networks," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2461–2470, 2022, doi: 10.1016/j.jksuci.2020.09.017.

[17] B. Vidya and P. Sasikumar, "Wearable multi-sensor data fusion approach for human activity recognition using machine learning algorithms," *SSRN Electronic Journal*, vol. 341, p. 113557, 2022, doi: 10.2139/ssrn.4014024.

[18] Y. Himeur, A. Alsalemi, A. Al-Kababji, F. Bensaali, and A. Amira, "Data fusion strategies for energy efficiency in buildings: overview, challenges and novel orientations," *Information Fusion*, vol. 64, pp. 99–120, 2020, doi: 10.1016/j.inffus.2020.07.003.

[19] R. Lavanya and N. Shanmugapriya, "An intelligent data fusion technique for improving the data transmission rate in wireless sensor networks," *International Journal of Computational Intelligence and Applications*, vol. 22, no. 01, Mar. 2023, doi: 10.1142/S1469026823410043.

[20] U. K. Lilhore *et al.*, "A depth-controlled and energy-efficient routing protocol for underwater wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 18, no. 9, p. 155013292211171, Sep. 2022, doi: 10.1177/15501329221117118.

[21] S. Govindaraj, L. Raja, S. Velmurugan, and K. Vijayalakshmi, "Extreme learning machine optimized by artificial cell swarm optimization for the data fusion modal in WSNs," *Peer-to-Peer Networking and Applications*, vol. 17, no. 3, pp. 1344–1357, 2024, doi: 10.1007/s12083-024-01643-9.

[22] J. Zhang, Z. Lin, P. W. Tsai, and L. Xu, "Entropy-driven data aggregation method for energy-efficient wireless sensor networks," *Information Fusion*, vol. 56, pp. 103–113, 2020, doi: 10.1016/j.inffus.2019.10.008.

[23] X. Liu *et al.*, "Kalman Filter-based data fusion of Wi-Fi RTT and PDR for indoor localization," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8479–8490, 2021, doi: 10.1109/JSEN.2021.3050456.

[24] R. Dautov, S. Distefano, and R. Buyya, "Hierarchical data fusion for smart healthcare," *Journal of Big Data*, vol. 6, no. 1, p. 19, Dec. 2019, doi: 10.1186/s40537-019-0183-6.

[25] S. S. Saranya and N. S. Fatima, "IoT-based patient health data using improved context-aware data fusion and enhanced recursive feature elimination model," *IEEE Access*, vol. 10, pp. 128318–128335, 2022, doi: 10.1109/ACCESS.2022.3226583.

BIOGRAPHIES OF AUTHORS

Mr. Naganna Shankar Sollapure is a diligent techno savvy assistant professor with excellent academic track record having 20 years of teaching experience in the Department of Electronics and Communication Engineering. He is pursuing full time Ph.D. from Visvesvaraya Technological University, Belagavi under QIP programme in the field of wireless sensor network. He carries his academic credentials with Bachelors of Engineering in Electronics and Communication Engineering (2000) from Karnatak University Dharwad and Master of Technology in Digital Systems and Communication Engineering (2007) from National Institute of Technology (Deemed), Calicut. He rendered his service at various levels in teaching, initially as lecturer, then assistant professor and he has been into research for the last few years. He has attended several seminars, national and international conferences, symposiums and workshops. He is a life member of ISTE. His fields of interests are: signal processing, speech processing, digital systems, wireless communication, and sensor networks. He can be contacted at email: naganna.sollapure@gmail.com.

Dr. Poornima Govindaswamy a committed academician with over 24 years of experience in Department of Electronics and Communication, B.M.S. College of Engineering. She obtained her Bachelor of Engineering in Electronics and Communication Engineering from Bangalore Institute of Technology and received her Master's degree in Digital Communication Engineering from BMS College of Engineering. She holds a Ph.D. in the energy efficient and fault tolerant wireless sensor networks from University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, Karnataka, India. Her research interest includes energy efficient computer networks, fault tolerant signal processing, and energy harvesting in wireless sensor networks. She has an extensive experience in designing curriculum, developing education pedagogy. She has published over 30 research papers and edited course materials. She has developed courses for VTU EDUSAT and e-Shikshana Programmes. She has fetched grants from MHRD, DRDO, Bengaluru. She was the Organizing Chair of the IEEE First International Conference on Networking Embedded and Wireless Systems-2018. She is in the editorial board of Glacier Journal of Scientific Research. She has severed as an evaluator/advisor in several evaluation committees constituted by UPSC, AICTE, KPSC, VTU. She is an IEEE Senior Grade Member, Fellow IETE and a Life Member of ISTE. She was awarded with Adarsh Vidya Saraswati Rashtriya Puraska in 2018 and Excellence in Higher Education in the year 2019 from the Center of Leadership Development, Venus International Foundation. She can be contacted at email: gpoornima.ece@bmsce.ac.in.