無線センサネットワークにおけるAWDFとLSTMを用いた水環境モニタリングのデータ融合技術の研究

Master's Thesis

STUDY ON DATA FUSION METHOD BASED ON AWDF AND LSTM FOR WATER ENVIRONMENT MONITORING IN WSNs

September 2023

Graduate School of Marine Science and Technology

Tokyo University of Marine Science and Technology

Master's Course of Marine System Engineering

SHEN BIN

Master's Thesis

STUDY ON DATA FUSION METHOD BASED ON AWDF AND LSTM FOR WATER ENVIRONMENT MONITORING IN WSNs

September 2023

Graduate School of Marine Science and Technology

Tokyo University of Marine Science and Technology

Master's Course of Marine System Engineering

SHEN BIN

Content

ABSTRACT

Water is the source of life, and a good water environment is even more vital for our life and work. To closely cooperate with the policy of protecting water resources and building a good water environment for living and working, the technology about water environment monitoring research is gradually developing in recent years. This combines the communication technology, computer technology and the Internet of Things (IoT) in one of the Wireless Sensor Networks (WSNs) of water environment monitoring system, because of its wide distribution, the formation of a flexible network, real time sensing and many other advantages and convenience, more suitable for the complex and changing requirements of the water environment monitoring. However, most of the water pollution detecting results are often relatively partial, and there are large deviations from the actual water quality, and this will cause a large degree of impact on the subsequent assessment of the water environment quality, prediction, and other processing. At the same time, a large amount of raw data directly transmitted to the monitoring and processing center will inevitably increase the amount of data transmission of the node, which in turn leads to network congestion and increased energy consumption. Therefore, it is necessary to use data fusion technology to process various water quality parameters obtained from WSNs.

In this article, water environment monitoring based on WSNs is used as a research background, and an industrial water purification plant in Chiba prefecture is selected as a specific data collection site. In view of the above problems, an area specific and feasible data fusion method for water environment intelligent monitoring system is proposed. Combined with the actual situation and characteristics of water quality monitoring in local water purification plants, pH, water temperature, turbidity, chromaticity, electrical conductivity, and other indexes of industrial water are selected as the measurement parameters of water environment monitoring, and these data are fused and processed. A multi-level data fusion method is proposed for the specific situation of the water purification plant. At the Data Level, Adaptive Weighted Data Fusion (AWDF) methods are used to initially process the

raw data and reduce the amount of data transmitted. And for the original algorithm, the weight coefficients are difficult to determine, the accuracy is insufficient and other problems, to optimize and improve it. At the Feature Level, a neural network-based data fusion algorithm is used, using multiple measurement parameters as input to the network, and the water environment is classified through sample training and preliminary judgment of the current water quality situation. Finally, at the Decision level, an optimization-based LSTM deep neural network is used to further predict future changes in single water quality to multi-parameter water quality.

The simulation results show that the multi-level data fusion method proposed in this article can monitor the water environment more comprehensively and accurately, discriminate the current water quality more effectively, and predict the water quality parameters between a period. These can provide favorable basic support for the subsequent strategy of water environment monitoring.

KEY WORDS: water environment monitoring, multi-level data fusion, AWDF, neural network, LSTM

1. Introduction

1.1 Research background and significance

To protect water resources, Wireless Sensor Networks (WSNs) are being widely used for water quality monitoring in different areas[1]. Instead of manually measuring the water quality parameters, the inspectors can realize the automatic collection of data by using the sensor nodes deployed in the target waters, and the collected data can be transmitted to the aggregation node through the sensor nodes in different areas, from which the water quality parameters of the whole monitoring waters can be further obtained, thus monitoring the water environment in a more flexible way[2]. However, in this process, the amount of environmental data collected by the sensor nodes is huge and easily interfered by external conditions, if the acquired raw data is not processed[3],on the one hand, it is likely to generate a large amount of redundant and abnormal data, increasing the energy loss in the transmission process .On the other hand, it is not conducive for the monitor to analyze the large amount of raw data acquired by the nodes, thus affecting the assessment and prediction of water quality. Therefore, processing (data fusion) of water quality data acquired by monitoring nodes is necessary. For example, abnormal and duplicate data can be removed to obtain a more comprehensive data from many original data. The correlation between multiple environmental parameters is used to comprehensively judge whether an indicator in the monitoring area meets the requirements. Through the collected data, analysis and prediction of future data and its changes, these are the basic contents of data fusion.For different monitoring environments, selecting multiple data fusion methods according to different needs can not only alleviate the transmission pressure of the network, extend the life cycle of the network[4], but also provide more intuitive and effective data evaluation and prediction methods for the monitoring personnel[5].

Taking the research of Zhang Xunuo, Zhao Ying et al on the water ecological function zoning of Songhua River Basin[6] as an example. The authors took the Songhua River basin

as the monitoring object, and took a variety of water quality environment and biological environment parameters as the original data. Through data fusion technology, according to different regional characteristics (for example, the ecological environment of the first region is good, the agricultural area of the second region is large, and the river network density of the third region is high), The ecological functions of the target watershed were divided (1-4) and the importance of the respective zones was assessed. This has promoted the protection of water resources, the maintenance of biodiversity and the development of agriculture in the region.

1.2 Main research content and purpose

This article focuses on water environment monitoring in WSNs and selects a portion of water quality parameters (such as water temperature, pH, turbidity, color, conductivity, etc.) from industrial water purification plants in Chiba prefecture between 2018 and 2022 as the data source. To address the two major difficulties in water environment monitoring: to determine whether the collected raw data are correct (whether there are anomalies, erroneous data, etc.) and whether the fusion accuracy meets the requirements; and whether it is possible to reasonably assess the current water quality situation as well as the changes in the water quality over a period under continuous observation. A suitable and feasible data fusion method for water environment monitoring is designed to fuse the acquired water quality parameters at three levels: data level, feature level and decision level.

Firstly, at the Data Level, an optimization scheme is proposed for the problems of the original adaptive weighted fusion algorithm to make it better applicable to the monitoring of the water environment. Then the initial numerical fusion is performed by MATLAB, and the original data without data fusion is compared and analyzed to further make the data more accurate for subsequent use.

Then, at the Feature Level, various water quality parameters (water temperature, pH, turbidity, conductivity, etc.) obtained from the water environment are extracted, and the data

fusion is determined at the feature level using an optimized neural network, which further gives the specific condition of the current water quality (water quality evaluation) for the next solution.

Finally, at the Decision Level, the LSTM deep neural network is designed and constructed to predict the water quality changes in the future period based on some time series parameters in the water environment obtained previously[7]. Specifically divided into water quality single-parameter prediction and multi-parameter prediction considering the correlation between multiple water quality, analyze and discuss the effect of these two time-series prediction methods and adaptation scenarios.

1.3 Structure of the article

Section 1 introduces the research background and significance of this article, the content and purpose of this study. Section 2 introduces the status of related research. Section 3 introduces the framework of the multi-level data fusion method proposed in this study. Section 4 illustrates the data fusion method of AWDF applicable to water environment. Section 5 discusses the data fusion mechanism for water quality evaluation based on neural networks. Section 6 presents the method of water quality parameter prediction using LSTM deep neural network. Section 7 evaluates the proposed multi-level data fusion method. Finally, Section 8 concludes this article.

2. Related works

2.1 Research status of water environment monitoring based on WSNs

WSNs with many features such as low cost, automation, distributed, high accuracy and spatio-temporal continuity have been successfully applied in various scenarios related to water environment monitoring, and many international universities and research institutions have conducted research on them, such as the University of California at Berwick, the University of São Paulo, Brazil, and the University of Bologna, Italy, have explored the solutions of WSNs in many different scenarios such as river snow storage measurement, mountain hydrological parameter measurement, and groundwater environmental parameter monitoring[8][9].

A distributed wireless sensor network for spatially scaled "water balance" monitoring was designed and deployed in a 2000 km^2 snow-dominated area in the upper American River Basin, California by a team from the University of California at Berwick. Welch et al. analyzed 11 years of historical data from the target area and used a clustering approach to determine the location of the sensor nodes. They found through their study that the monitoring performance of the network can be significantly improved by placing the right number of sensors.

Between 2012 and 2015, a multi-scale integrated observation experiment HiWATER was conducted in the Heihe River Basin, China[10], with the interplay of satellite and airborne remote sensing and ground-based observations. the experiment significantly improved the observation capability of the water environment and established a leading hydrological observation system.

2.2 Research status of data fusion technology

In 2019, Sun Guiling, Zhang Ziyang[11] published a study on the use of multiple sensors for data fusion in greenhouse environments. To address issues such as low fusion accuracy and poor interference resistance, they proposed a multi-sensor data fusion algorithm based on trust and an improved genetic algorithm. The raw data collected by the sensors is transmitted to the gateway through receiving nodes, where data preprocessing is performed to eliminate abnormal data. Then, using fuzzy theory, the preprocessed data is subjected to trust-based fusion operations, avoiding absolute trust among the data. Finally, an improved genetic algorithm is employed to optimize the fused estimation values. Experimental results demonstrate that this fusion method can ensure accuracy while reducing the execution time of the algorithm, making it a feasible data fusion approach.

In 2022, Gong Li, Yan Jinlong[12][12] published a paper on an Internet of Things (IoT) intelligent irrigation system with data fusion capabilities. This study optimized irrigation plans by incorporating data fusion techniques into a conventional irrigation system. It employed LSTM (Long Short-Term Memory) and integrated diverse data sources such as historical weather data, past irrigation logs, weather forecasts, and Wireless Sensor Networks (WSNs). By simulating and predicting irrigation requirements, the system achieved significant improvements in water Water-saving efficiency. efficiency compared to traditional automatic timer-based irrigation methods.

3. Overview of data fusion methods based on the water environment

3.1 Concept and significance of data fusion

Data fusion is an information processing method that uses computers to automatically analyze several measurement data obtained in a temporal sequence under certain guidelines to accomplish the required decision-making and evaluation tasks. The sensors collecting the observed data have limited capacity in storage, data forwarding and data calculation, etc. If these data are directly utilized without processing, it will not only cause more energy loss in the process of data transmission, but also lead to many problems such as lack of data accuracy and inaccuracy. For the final monitoring results will also have a large deviation from the actual situation[13], thus affecting the assessment of the monitoring area. Therefore, it is necessary to perform data fusion on the raw data collected by the sensors to reduce duplicate and abnormal data, reduce energy consumption, and evaluate the situation of the monitoring area more comprehensively[14].

As shown in Fig.3.1, for multiple different environmental parameters a, b, c, d, and e, the spatial position fusion of the same sensor (1-n represents the data of sensor nodes at different locations) can be used to obtain a larger range of unique fusion values from many raw data collected in a small range. For the time series fusion of the same sensor (where 1-n represents the time series data of the same parameter), the change trend of the data in a certain time can be intuitively reflected. The abnormal data can be eliminated, and its weight reduced in the fusion process to reduce the impact on the fusion result.

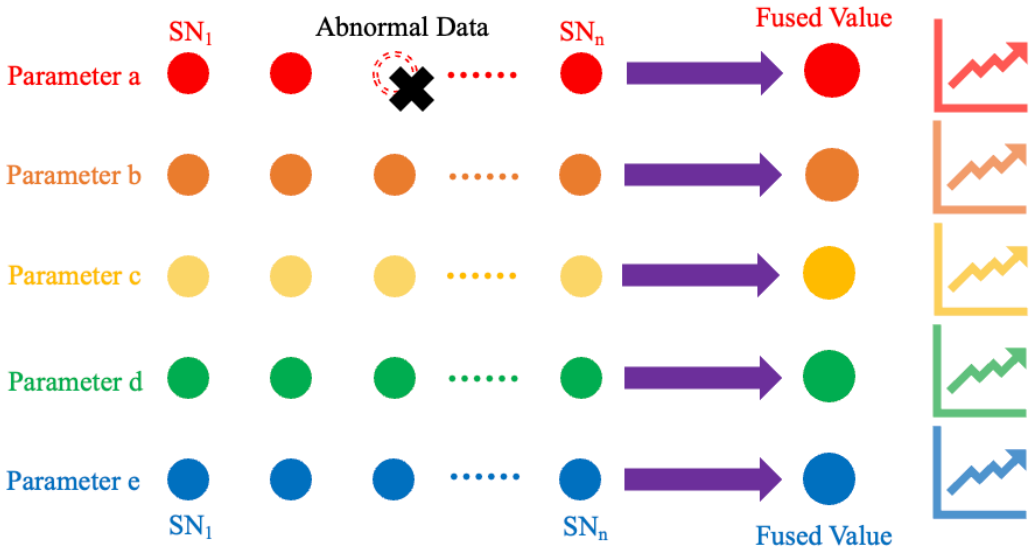

Fig.3.1 Data fusion of single sensor

In practical application scenarios, there is often more than one type of sensor involved in data collection, but many different types of sensors working together, each collecting different environmental parameters[15]. Each type of sensor has its own characteristics, and a specific sensor can only collect the required parameters in a specific range, which requires a method that can use the correlation between multiple information for comprehensive processing and evaluation, which is multi-sensor data fusion technology.

The spatial data fusion of multiple sensors is shown in Fig.3.2. The leftmost region represents the divided monitoring sub-region (1-n), different colors represent different sensor nodes, and the fusion value in the middle is the fusion value of each sub-region (this fusion value can be any known environmental parameters or new parameters associated with these environmental parameters). Using the single sensor spatial fusion method in Fig.3.1, The final fusion value of the entire monitoring area can be obtained.

The time data fusion of multiple sensors is shown in Fig.3.3. The leftmost part represents a time node (1-n), which contains multiple environmental parameters collected at this moment, and the fusion value in the middle is the fusion value of the current moment (this fusion value can be any known environmental parameters, or it can be a new parameter associated with these environmental parameters). Using the single sensor time fusion method

in Fig 3.1, we can know the time series changes of the target parameters in the monitoring area.

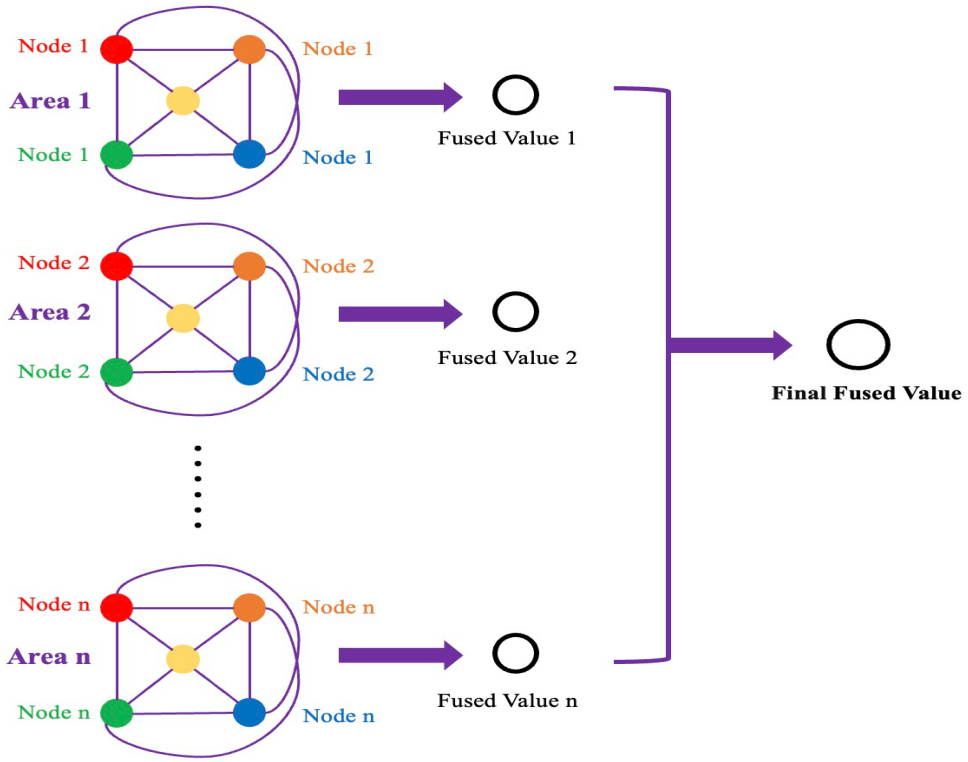

Fig.3.2 Spatial data fusion of multi sensor

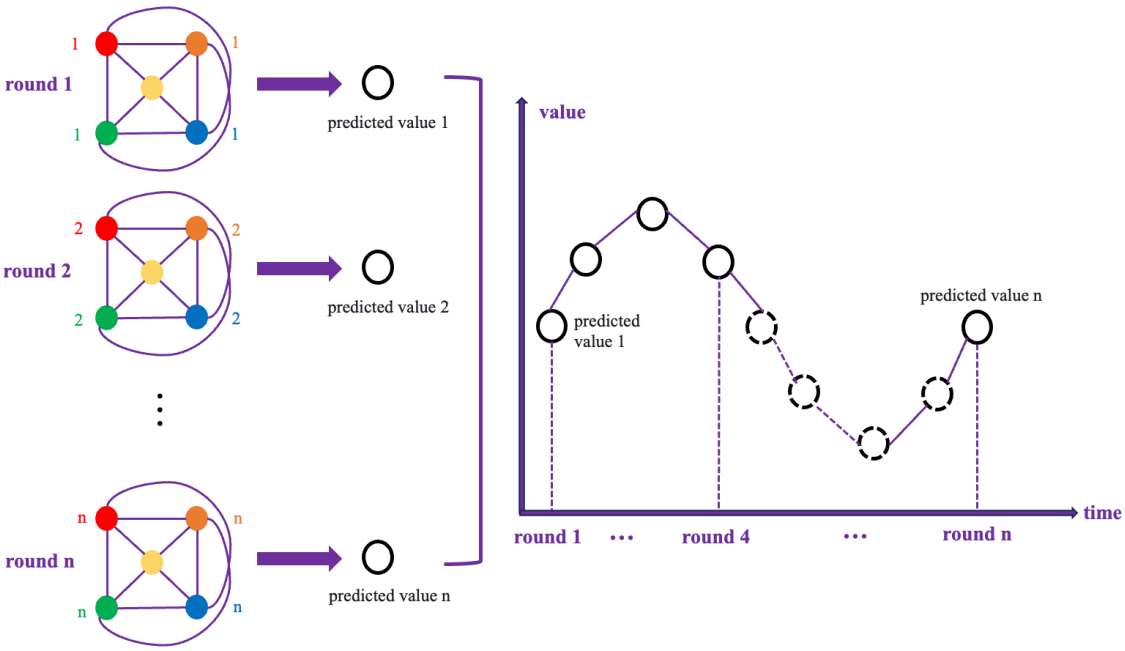

Fig.3.3 Time data fusion of multi sensor

The multi sensor data fusion technique Compared with single sensor data fusion, multi-sensor data fusion can collect environmental parameters more comprehensively and accurately, analyze the possible internal correlations between multiple data, and thus make more reasonable judgments and decisions.

3.2 Classification and methods of data fusion

3.2.1 Classification of data fusion

According to the processing type of data fusion, the variation of data volume and the level of fusion, as shown in Fig.3.4. we can classify data fusion into the following different types[16].

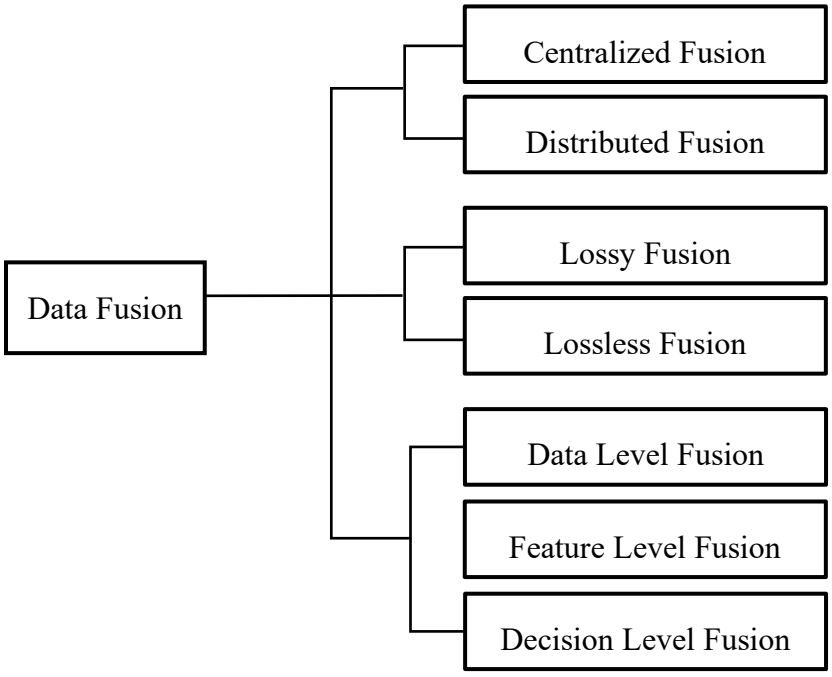

Fig.3.4 Classification of data fusion

(1) Classification by the type of data fusion structure

Centralized Fusion: the data collected by all sub-nodes are transmitted directly to the aggregation node without any processing, which will greatly increase the channel pressure and may cause information blocking, but on the other hand, this approach preserves the

integrity of the data to the maximum extent.

Distributed Fusion: in contrast to centralized, the data collected by sub-nodes are processed before transmission, which effectively reduces the amount of data transmission and extends the effective working time of nodes. However, it is easy to cause data loss and accuracy is difficult to guarantee.

(2) Classification according to the number of variations of data fusion

Lossy Fusion: save energy of each link by reducing the amount of data transmitted, but it must ensure that the remaining data contains valid information, otherwise the final fusion result will be difficult to guarantee.

Lossless Fusion: processing based on ensuring data integrity and only roughly grouping part of the data, which can reduce the duplicated grouped header information, but the fusion effect is often not ideal because the operation is too simple.

(3) Classification according to the level of data fusion

Data Level Fusion: it is a lower level of fusion, in which similar data acquired by nodes in the monitoring area are fused. Therefore, this method is usually only applicable to the case of single environmental parameter acquisition and cannot be adapted to the fusion of multiple parameters in complex environments. However, it can still be used as an initial fusion method for the initial processing of the raw data numerically. The adaptation to multi-parameter data fusion needs to be further improved by combining other methods in specific application scenarios.

Feature Level Fusion: this method belongs to the intermediate level, where the multivariate data collected from the nodes of the monitoring area are abstracted to obtain feature values to obtain possible correlations between multiple data, which in turn can better fuse the original data and provide evaluation criteria for subsequent assessments.

Decision Level Fusion: it is more advanced compared to the first two data fusion methods, which can not only extract correlations between various parameters, but also provide future

decisions by reacting to changes in the detection area over a period based on changes in the time series.

3.2.2 Data fusion methods

They can be broadly divided into two categories: classical methods and modern methods. Among them, classical methods are subdivided into estimation and statistical methods. Modern methods are divided into information theory and artificial intelligence methods. A more detailed classification is shown in the Fig.3.5.

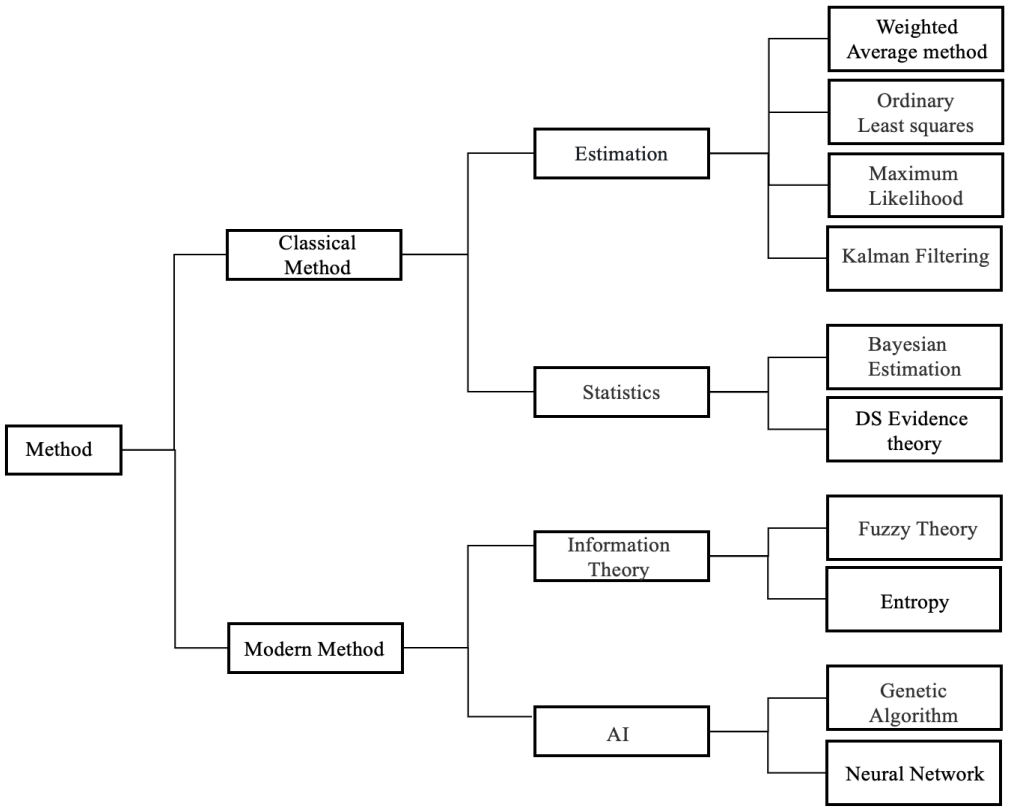

Fig.3.5 Common data fusion methods

3.3 Analysis and design of data fusion in water environment monitoring

The sensor nodes for the measuring water environment are randomly distributed on the water, and the raw data received by the sink nodes in small-scale areas are prone to redundancy, anomalies, etc. If data fusion is not used for its weighting of the appropriate increase or decrease, then it is difficult to exclude the interference of the sensors themselves and the outside of the environment. In addition, because the environmental parameters are diverse, it is important to establish correlation between multiple heterogeneous data, through the input of multiple parameters to determine the current regional water quality. Finally, using the historical data of environmental parameters already obtained, further predict the data changes over a period.

Based on the above analysis and in combination with the specific requirements for water quality monitoring in industrial water treatment plants, this research proposes a multi-level data fusion method. The specific processing workflow is illustrated in Fig.3.6. Firstly, the target water environment under monitoring is taken as the research object, and various types of sensors (such as temperature and pH sensor) are used to measure the required environmental parameters. These parameters are then used as the data fusion processing objects and are subjected to three levels of data fusion separately. In the data level, a refined Adaptive Weighted Data Fusion[17] (AWDF) algorithm is utilized for a single parameter (temperature) obtained from the same type of sensor. This is done to address issues like the inability to place many sensors in the actual environment and the presence of anomalous data. The AWDF algorithm increases the fusion frequency and reduces the weight of anomalous data to further enhance the fusion accuracy. In the feature level, a Backpropagation Neural Network (BPNN) is employed to analyze the fusion of multiple parameters (temperature, pH, turbidity, color, conductivity) obtained from different sensors. This process provides a comprehensive assessment of the water quality in the monitored area, enabling the implementation of further measures. At the decision level, historical data

collected from various parameters are used in conjunction with Long Short-Term Memory (LSTM) to make single parameter predictions (based solely on the historical data of individual parameters) and multi parameter predictions (analyzing historical data from multiple parameters together) for a certain period. This approach facilitates better monitoring and management of industrial water quality.

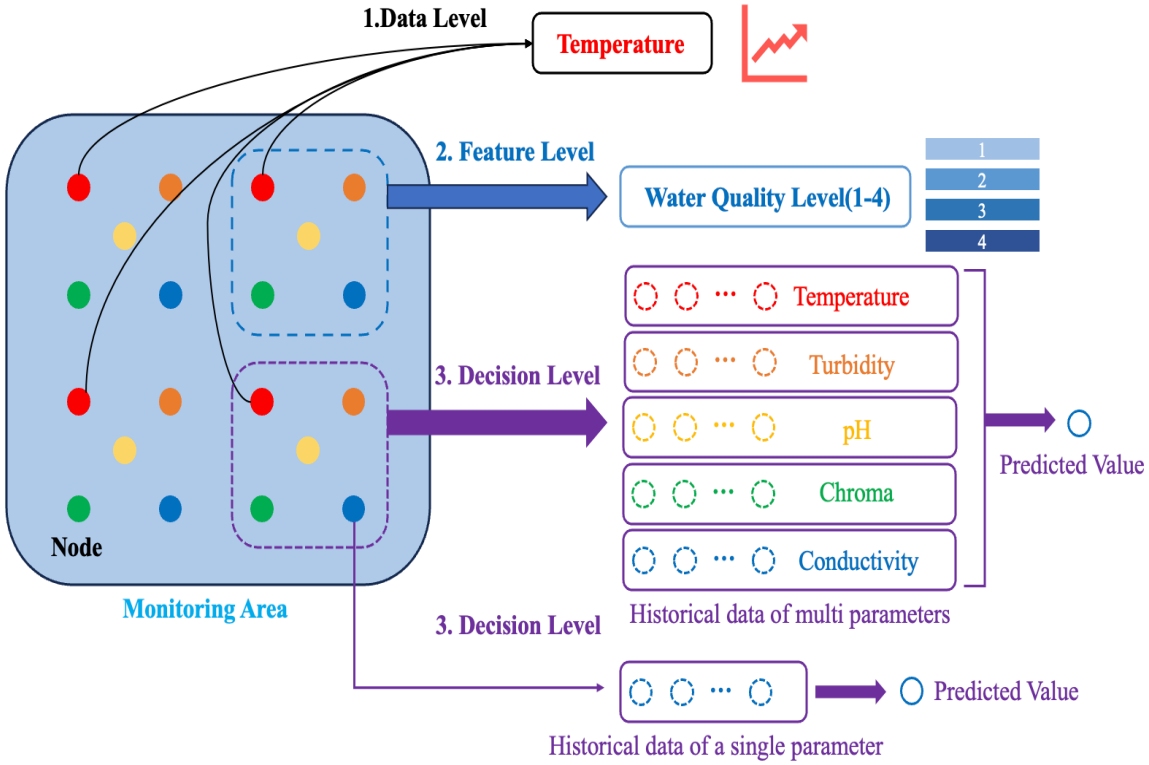

Fig.3.6 Three-level data fusion mechanism

4. Data fusion method based on AWDF

4.1 Principle analysis of AWDF

In an actual area to be measured, usually multiple sensors in different directions for its data acquisition and processing, and then get the environmental information in the area. Compared to a single sensor, a more accurate estimate will be obtained, but there will still be other interfering factors, such as interference from the external environment, differences in the individual sensors themselves, etc. How to solve this problem becomes the key to the first step of data-level fusion.

Suppose there are n sensor nodes to collect and record the results of the parameters in the water environment, respectively $X_1, X_2, \ldots, X_n$ under the premise of ensuring the minimum total variance of the parameter using the adaptive weighting fusion algorithm (AWDF), for each of these nodes to collect the data are assigned the corresponding weighting coefficient W_i, when the weighting coefficient is optimal[18], the final fusion result $\hat{X}$, the corresponding fusion structure of this algorithm is shown in Fig. 4.1.

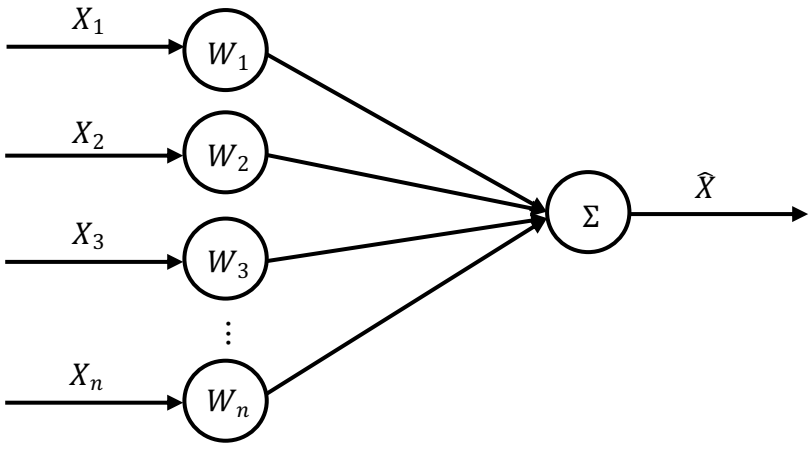

Fig.4.1 Model of AWDF algorithm

This method gives a different coefficient W_i between 0 and 1 for each sample data, it represents the reliability of the data. The closer to 1 means that this data is more reliable, and

the opposite is less reliable. However, no arbitrary values can be assigned to the node data according to subjective wishes, and they can only be derived by ensuring that the total variance is minimal.

For acquisition data $X_1, X_2, \dots, X_n$, there are also n corresponding weight coefficients $W_1, W_2, \dots, W_n$. The relationship satisfies the following equation (4-1).

$$\begin{cases} \hat{X} = \sum_{i=1}^{n} W_i X_i \\ \sum_{i=1}^{n} W_i = 1 \end{cases} \tag{4-1}$$

The total mean square error σ^2 of the response according to the theoretical derivation is:

$$\sigma^2 = E\left[\left(X - \hat{X}_i\right)^2\right]$$

$$= E\left[\sum_{i=1}^{n} W_i^2 (X - X_i)^2 + 2 \sum_{i=1, j=1, i \neq j}^{n} W_i W_j (X - X_i)\left(X - X_j\right)\right] \tag{4-2}$$

Since the n data $X_1, X_2, \dots, X_n$ are initially independent of each other and are also unbiased estimates of X, the following equation holds.

$$E(X - X_i)\left(X - X_j\right) = 0 \quad (i = 1,2, \dots, n; j = 1,2, \dots, n; i \neq j) \tag{4-3}$$

It can be further known that its total mean square error σ^2 can be expressed as:

$$\sigma^2 = \sum_{i=1}^{n} W_i^2 \sigma_i^2 \tag{4-4}$$

where σ_i^2 denotes the mean square error of each node.

Through mathematical knowledge we can know that there is a minimum value of the total mean square error σ^2 in equation (4-4), and the required value can be solved using equation (4-5) below.

$$\begin{cases} \sigma_{min}^2 = \min\left(\sum_{i=1}^{n} W_i^2 \, \sigma_i^2\right) \\ \sum_{i}^{n} W_i = 1 \end{cases} \tag{4-5}$$

The auxiliary function θ is constructed according to the theory of conditional extrema of

multivariate functions.

$$\theta(W_1, W_2, \dots, W_n, \lambda) = \sum_{i=1}^{n} W_i^2 \sigma_i^2 - \lambda\left(\sum_{i=1}^{n} W_i = 1\right) \tag{4-6}$$

Construct the corresponding set of equations：

$$\begin{cases} \dfrac{\partial\theta}{\partial W_1} = 2W_1\sigma_1^2 - \lambda = 0 \\ \dfrac{\partial\theta}{\partial W_2} = 2W_2\sigma_2^2 - \lambda = 0 \\ \cdots \\ \dfrac{\partial\theta}{\partial W_n} = 2W_n\sigma_n^2 - \lambda = 0 \\ \dfrac{\partial\theta}{\partial\lambda} = 1 - \sum_{i=1}^{n} W_i = 0 \end{cases} \tag{4-7}$$

Using the above equation (4-7), the optimal weight coefficient for each node is calculated when the minimum value of the total mean square error σ^2 in equation (4-5) is obtained as:

$$W_i = \frac{1}{\sigma_i^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2}} (i = 1,2, \dots n) \tag{4-8}$$

The minimum value of the corresponding total mean square error σ_{min}^2 is：

$$\sigma_{min}^2 = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \tag{4-9}$$

In this way, the final fusion result can be found by using equation (4-1).

4.2 AWDF for water environment

4.2.1 Algorithm improvement

According to the derivation of the formula in the previous section, it is easy to find that the accuracy of the final fusion value depends largely on the number of nodes and the number of times each data is processed and fused. Considering the actual water environment monitoring situation, the common AWDF needs to wait for each group of data to arrive before the unified processing, which may reduce the efficiency of the network and increase the delay. To solve these problems and improve the fusion accuracy, this article makes an

optimization for the common AWDF algorithm and proposes a new superposition virtual AWDF algorithm (SV-AWDF). The basic idea is to increase the number of fusions, reduce the waiting time and further improve the fusion accuracy by adding virtual sensor nodes. The approximate fusion is shown in Fig.4.2.

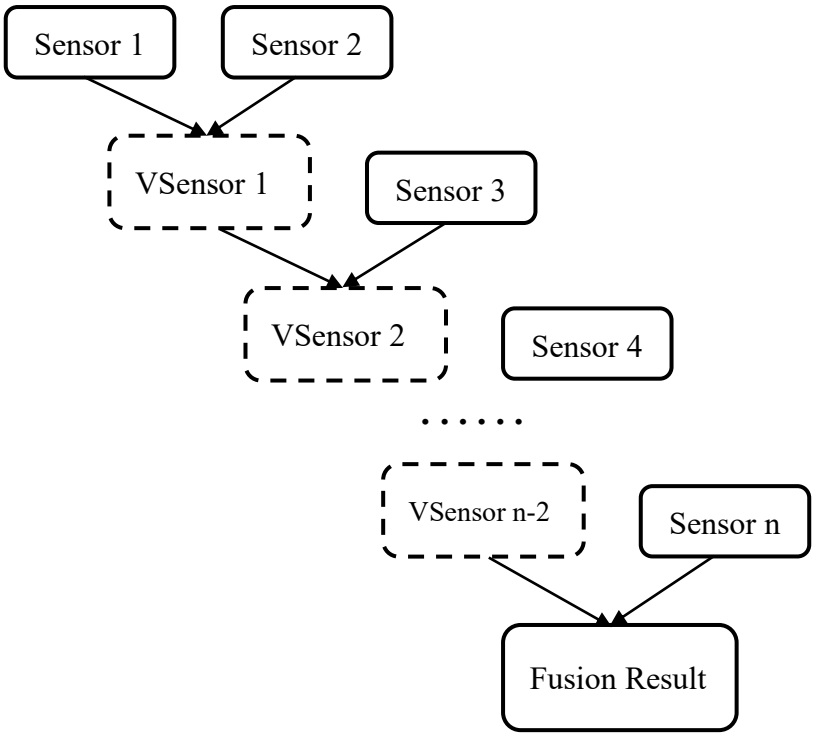

Fig.4.2 schematic diagram of the SV-AWDF

The specific convergence flow chart is shown in Fig.4.3.

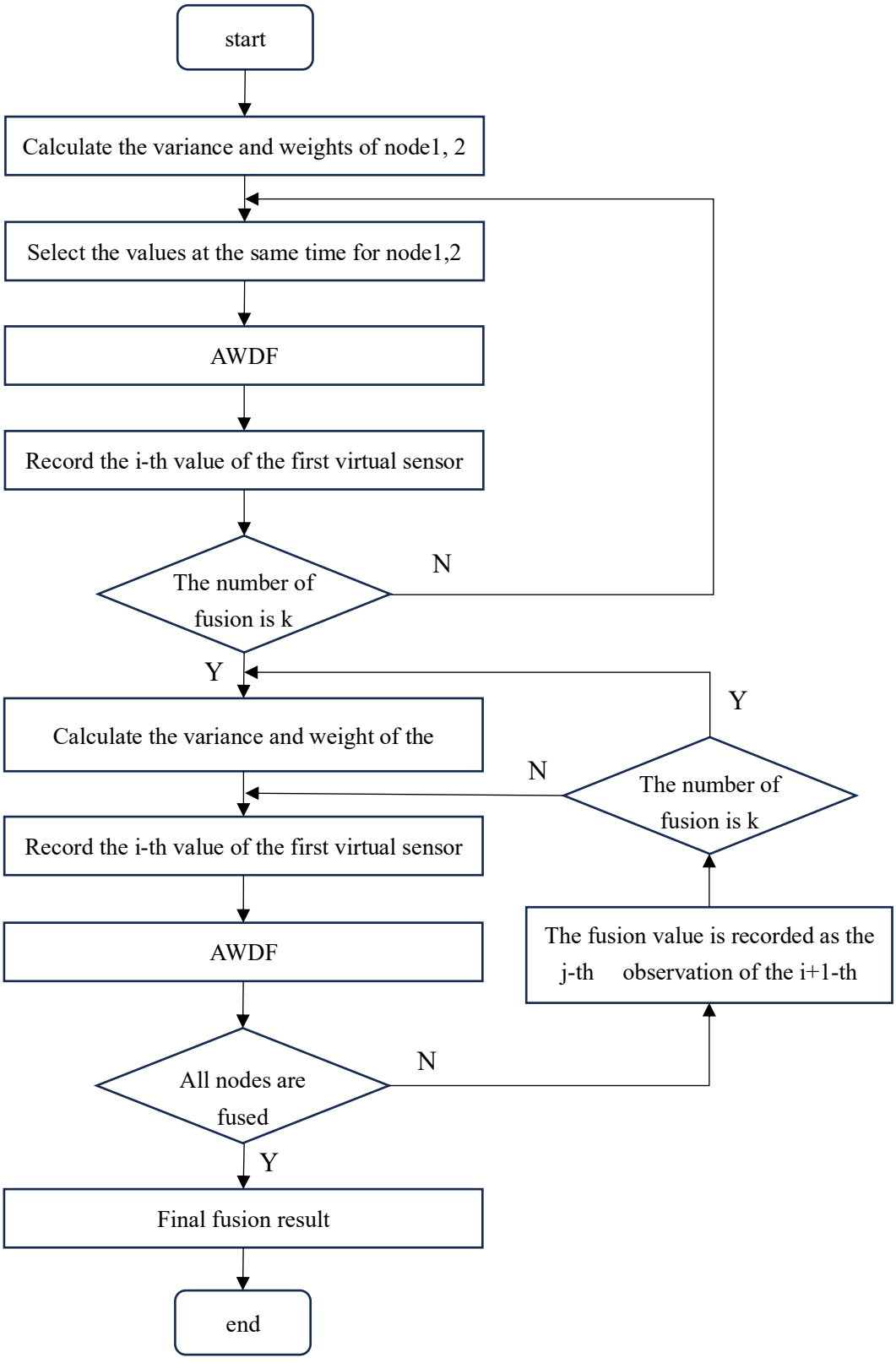

Fig.4.3 SV-AWDF work flow chart

4.2.2 Fusion results and performance analysis

Considering the actual monitoring situation in water environments, the original data collected from the water treatment plant, including parameters such as water temperature, pH, turbidity, color, and conductivity, were used as the acquisition node's raw data. The optimized AWDF algorithm was applied for the Data Level Fusion. The approximate process is shown in the Fig.4.4. $SN_1 - SN_n$ represents the number of sensors measuring the same parameter, and the data collected by all sensors at the same time is used as input to obtain the unique fusion value at the same time.

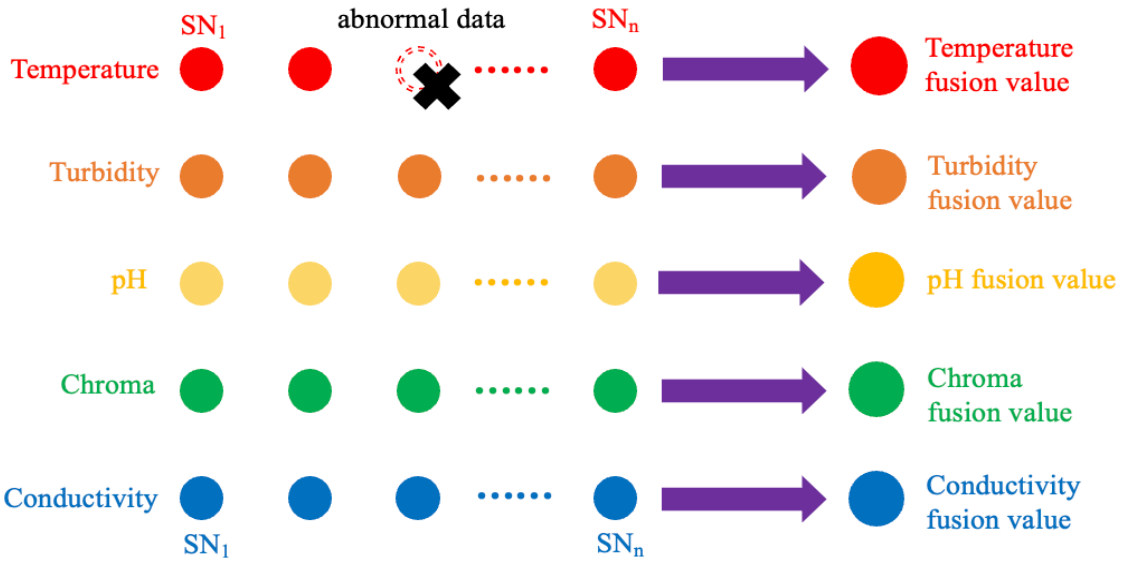

Fig.4.4 Data level fusion diagram

The following table 4.1 shows a portion of the collected raw data.

Table 4.1 Partial water environmental parameters

Temperature	pH	Turbidity	Chroma	Conductivity
16.5	7.8	26.0	31	20.4
15.5	8.1	22.7	29	20.9
16.0	8.1	19.8	26	21.2
18.0	8.9	19.7	25	22.7
19.0	9.0	22.8	26	23.0

21.0	9.2	18.9	21	21.7
20.0	9.3	21.6	25	23.1
19.5	9.4	24.0	28	21.4
19.0	9.3	24.1	29	23.3
18.0	8.9	26.1	28	24.6
19.5	9.3	25.9	26	23.5
21.0	9.5	20.3	22	22.5
20.5	9.2	22.6	29	23.9
20.0	8.3	25.5	34	22.4
19.0	8.3	29.9	38	21.6
18.5	7.9	21.2	31	21.8
17.0	7.7	15.1	23	22.1
19.5	8.3	18.9	26	21.5
20.5	8.8	20.8	25	22.5
22.0	9.1	24.3	27	22.5
21.5	9.2	19.2	22	22.8
20.0	9.1	24.5	30	23.2
21.5	9.4	32.2	34	22.6
21.0	9.2	29.3	32	23.1
22.0	8.8	30.2	34	24.5
23.5	9.2	31.7	34	22.8
20.5	8.9	42.2	48	24.6
24.5	9.1	31.0	35	27.1
24.5	9.1	30.7	33	27.1
23.5	9.1	36.3	40	27.1
20.5	9.0	45.8	54	24.4
21.5	8.1	26.5	38	29.0

23.5	9.1	39.2	46	26.6
23.0	9.0	26.2	35	24.6
22.5	8.9	32.9	40	23.9
23.0	8.9	36.5	45	23.6
19.0	8.6	55.1	70	24.9
19.5	8.4	32.0	47	24.9
18.5	8.5	37.4	49	24.3
18.5	8.6	35.1	42	24.4
21.0	9.1	34.5	44	24.4
21.0	8.6	24.5	32	25.4
21.0	8.7	34.3	40	25.1
21.0	8.5	29.1	37	25.2
19.5	8.6	49.2	67	23.8
19.0	8.7	40.6	48	24.6
20.0	8.9	32.7	39	24.4
21.5	8.7	30.3	37	27.8
22.5	9.0	17.8	26	27.1
24.5	9.0	25.0	31	27.4
23.5	8.5	37.6	42	28.2
24.0	9.0	33.3	39	23.9
23.0	8.9	45.1	57	22.0
24.0	8.8	42.2	49	23.7
24.5	8.9	38.5	53	26.7
26.5	8.6	34.3	43	27.9
28.3	8.9	38.6	48	29.4
28.5	9.1	38.0	47	28.6
29.0	9.2	28.9	39	27.8

27.0	8.6	35.6	45	28.1
29.5	9.1	25.7	37	27.4

(Reference: Chiba prefecture open data site, Water quality information[19])

As shown in Fig. 4.5, the temperature data in the above table is fused using the proposed SV-AWDF method to obtain the temperature trend before and after fusion.

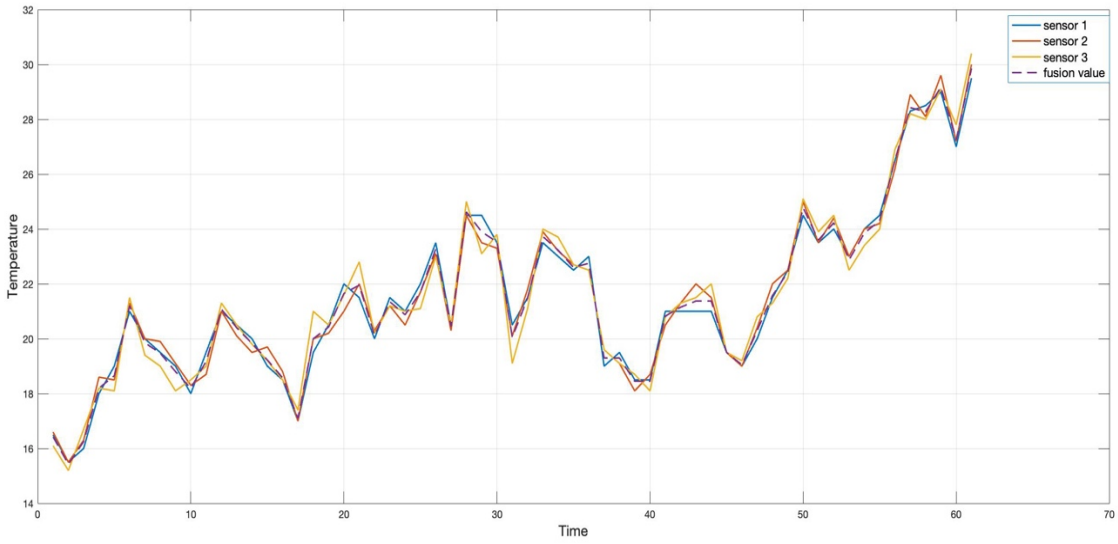

Fig.4.5 Temperature trends before and after fusion

The result of temperature change is shown in the Fig.4.6. Considering the lack of actual monitoring of possible abnormal data in the open data used, more obvious abnormal data is artificially added to the original data (the 12th data of sensor 3 is lower than that of the other two sensors, and the 17th data is higher than that of the other two sensors), and the SV-AWDF method is used again for data fusion processing. It is not difficult to find that this method can reduce the weight of abnormal data and give the final fusion result more accurately after adding abnormal data.

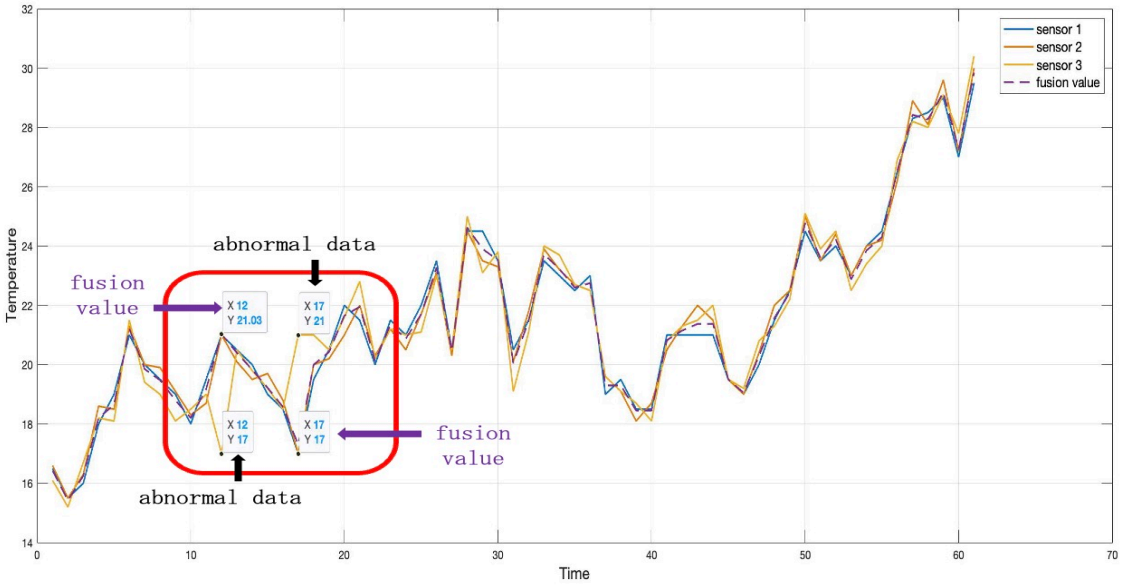

Fig.4.6 Temperature trends before and after fusion (add abnormal data)

In practical applications, even sensors of the same type may exhibit differences in data transmission, and the number of sensors placed in the water environment may be limited. To fuse the data more accurately, the proposed SV-AWDF method is utilized. If the first two sensors to transmit data are Node 1 and Node 2, the data from these two sensors are first subjected to regular AWDF fusion processing, resulting in the creation of the first virtual sensor. Then, this virtual sensor is combined with the next sensor through regular AWDF fusion, and the process continues until the last sensor is also included in the fusion process.

5. Data fusion method based on neural network

5.1 Introduction to Neural Network

In Feature Level data fusion, widely used techniques include the Dempster-Shafer theory of evidence and neural networks. In this study, the second-stage data fusion is performed using a backpropagation (BP) neural network[20][21], which has strong capabilities in parallel processing and data mapping. Its structure is shown in Fig.5.1.

The BP neural network possesses the following characteristics:

(1) The data storage form is decentralized, ensuring the stability of the network. It is not significantly affected by the interference of data from a specific node, thus minimizing the impact on the output.

(2) By relying on multidimensional mapping techniques, the BP neural network can achieve pattern recognition functionality and capture linear or nonlinear relationships in multi-input and multi-output scenarios.

Therefore, the BP neural network is suitable for the second-stage data fusion as it can effectively handle parallel processing and complex mapping relationships to enhance the fusion results.

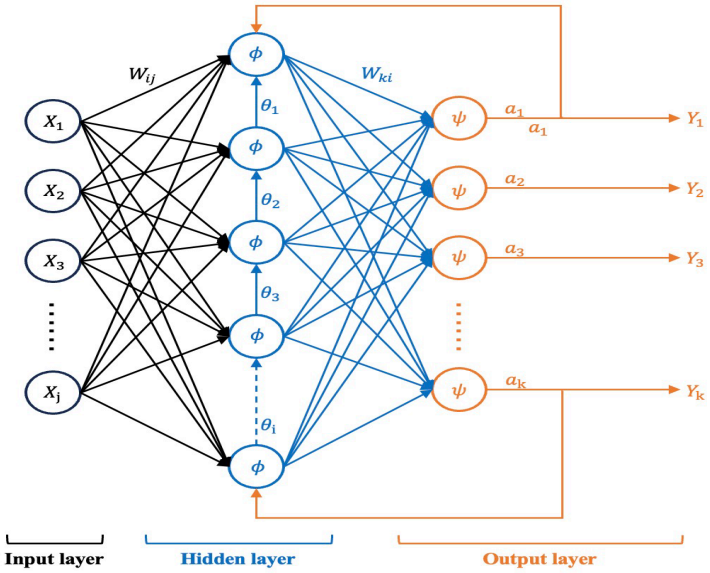

Fig.5.1 BP neural network structure

The explanations of each parameter in the above figure are as Table 5.1:

Table 5.1 Explanations of each parameter

parameter	explanation
X_j	The input to the j-th node in the input layer
W_{ij}	The weight between the j-th node in the input layer and the i-th node in the hidden layer
W_{ki}	The weight between the i-th node in the hidden layer and the k-th node in the output layer
θ_i	The threshold (bias) of the i-th node in the hidden layer
a_k	The threshold (bias) of the k-th node in the output layer
ϕ	The activation function of the hidden layer
ψ	The activation function of the output layer
Y_k	The output of the k-th node in the output layer

The specific algorithm is derived as follows.

(1) The forward propagation process of a signal:

Input net_i of the i-th node in the hidden layer：

$$net_i = \sum_{j=1}^{M} W_{ij} X_j + \theta_i \qquad (5-1)$$

Output $net_i$of the i-th node in the hidden layer：

$$Y_i = \phi(net_i) = \phi\left(\sum_{j=1}^{M} W_{ij} X_j + \theta_i\right) \qquad (5-2)$$

Input net_k the k-th node in the output layer：

$$net_k = \sum_{i=1}^{q} W_{ki} Y_i + a_k = \sum_{i=1}^{q} W_{ki} \phi \left(\sum_{j=1}^{M} W_{ki} X_j + \theta_i \right) + a_k \qquad (5-3)$$

Output net_k the k-th node in the output layer：

$$Y_k = \psi(net_i) = \sum_{i=1}^{q} W_{ki} Y_i + a_k = \psi \left[\sum_{i=1}^{q} W_{ki} \phi \left(\sum_{j=1}^{M} W_{ij} X_j + \theta_i \right) + a_k \right] \qquad (5-4)$$

(2) Back propagation of errors：

Reverse error propagation is a gradual adjustment from the output layer to the input layer using the error gradient descent method. The thresholds and weights of each layer of the network are continuously optimized, enabling the output of the desired results.

For each sample p the error criterion function E_P(T_k represents the expected output of the k-th node in the output layer.) ：

$$E_P = \frac{1}{2} \sum_{k=1}^{L} (T_k - Y_k)^2 \qquad (5-5)$$

The total error criterion function $E_p^{'}$ for p samples：

$$E_p' = \frac{1}{2} \sum_{p=1}^{p} \sum_{k=1}^{L} ({T_k}^p - {Y_k}^p)^2 \qquad (5-6)$$

According to the gradient descent method, the corresponding parameters of each layer are adjusted, which are the correction of the output layer weights ΔW_{ki}, the correction of the output layer thresholds $\Delta \alpha_k$, the correction of the implied layer weights ΔW_{ij}, and the correction of the implied layer thresholds $\Delta \theta_i$.

$$\Delta W_{ki} = -\eta \frac{\partial E}{\partial W_{ki}} \qquad (5-7)$$

$$\Delta \alpha_k = -\eta \frac{\partial E}{\partial \alpha_k} \qquad (5-8)$$

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} \qquad (5-9)$$

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i} \qquad (5-10)$$

Output layer weight adjustment：

$$\Delta W_{ki} = -\eta \frac{\partial E}{\partial W_{ki}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial W_{ki}} = -\eta \frac{\partial E}{\partial Y_k} \frac{\partial Y_k}{\partial net_k} \frac{\partial net_k}{\partial W_{ki}} \quad (5-11)$$

Output layer threshold adjustment：

$$\Delta \alpha_k = -\eta \frac{\partial E}{\partial \alpha_k} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial \alpha_k} = -\eta \frac{\partial E}{\partial Y_k} \frac{\partial Y_k}{\partial net_k} \frac{\partial net_k}{\partial \alpha_k} \quad (5-12)$$

Hidden layer weight adjustment：

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial W_{ij}} = -\eta \frac{\partial E}{\partial Y_k} \frac{\partial Y_k}{\partial net_k} \frac{\partial net_k}{\partial W_{ij}} \quad (5-13)$$

Hidden layer threshold adjustment：

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial \theta_i} = -\eta \frac{\partial E}{\partial Y_k} \frac{\partial Y_k}{\partial net_k} \frac{\partial net_k}{\partial \theta_i} \quad (5-14)$$

The final collation can be obtained:

$$\Delta W_{ki} = \eta \sum_{p=1}^{p} \sum_{k=1}^{L} ({T_k}^p - {Y_k}^p) \psi'(net_k) y_i \quad (5-15)$$

$$\Delta \alpha_k = \eta \sum_{p=1}^{p} \sum_{k=1}^{L} ({T_k}^p - {Y_k}^p) \psi'(net_k) \quad (5-16)$$

$$\Delta W_{ij} = \eta \sum_{p=1}^{p} \sum_{k=1}^{L} ({T_k}^p - {Y_k}^p) \psi'(net_k) \Delta W_{ki} \phi' net_i x_j \quad (5-17)$$

$$\Delta \theta_i = \eta \sum_{p=1}^{p} \sum_{k=1}^{L} ({T_k}^p - {Y_k}^p) \psi'(net_k) \Delta W_{ki} \phi' net_i \quad (5-18)$$

5.2 Design of NN in water environment monitoring

5.2.1 Process flow

In the second stage of feature level fusion, unlike data level fusion, this part deals with all types of sensors at the same time in the monitoring subregion. The general fusion process is shown in the Fig.5.2. In the monitoring sub-area, the measured values of the five sensors are taken as a group, and the values of these parameters are used to give the evaluation of the water quality class (1-4) at a certain moment in the current area.

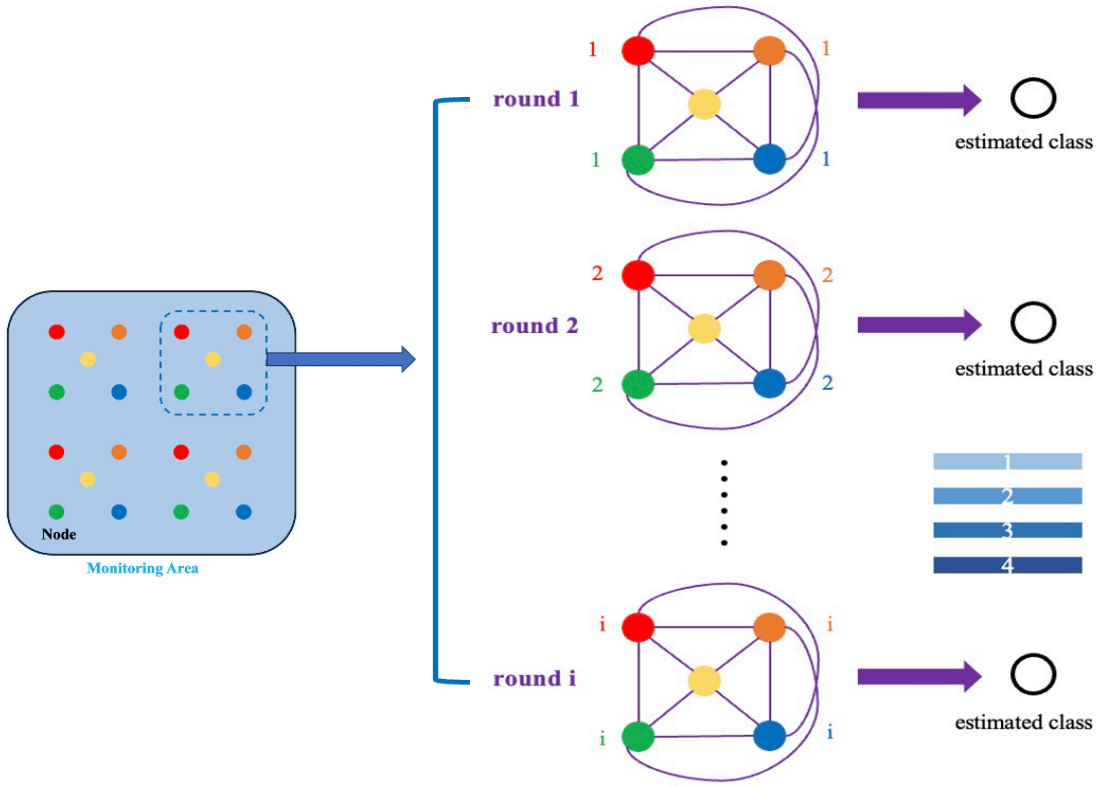

Fig.5.2 Feature level fusion diagram

In terms of BP neural networks, the process begins with the initial normalization of the entire network, where the weights and thresholds are initialized. Then, various environmental data collected by the water environment monitoring system is used as input parameters. Next, based on the input values from the samples, the input and output values of the nodes in the intermediate hidden layer are calculated. The next step involves calculating the input and output values of the nodes in the output layer. Finally, the output

layer error is computed, and the network weights are continuously adjusted to minimize the error, repeating the above steps to ensure that the overall mean square error meets the requirements. The Fig.5.3 shows the processing flow chart of BP neural network.

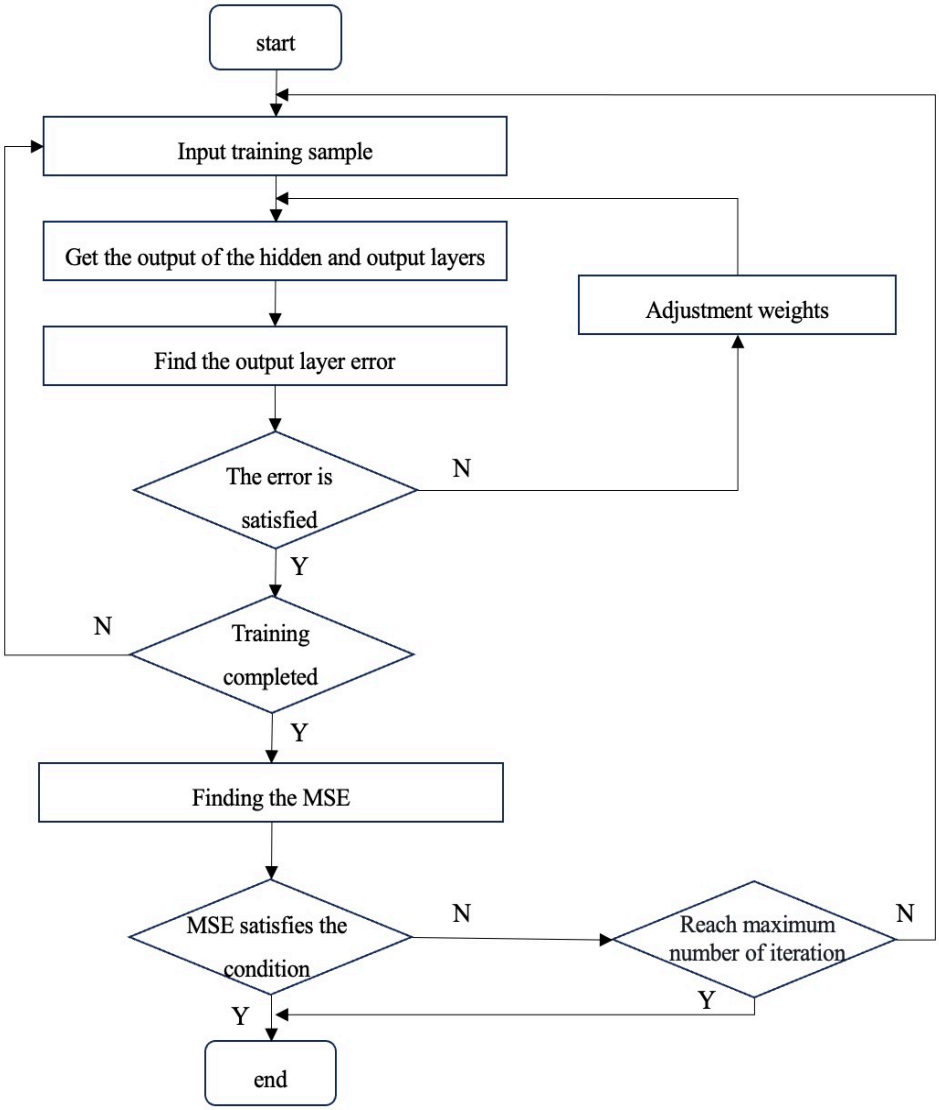

Fig.5.3 BP neural network work flow chart

5.2.2 Setting of related parameters

Five sample parameters of temperature, pH, turbidity, chroma, and conductivity in water quality testing in water purification plants were selected as the number of input neurons for this network, and the output layer was the assessment level of water quality (from good level 1 to poor quality level 4).

Firstly, the initial normalization is carried out for the whole network, as well as initializing the weights and thresholds of the network, and then a variety of environmental data collected in the water environment monitoring system is used as the input parameters, and then the input domain output values of the nodes in the middle implicit layer are calculated according to the sample input values. The next step calculates the input and output values of the nodes in the right output layer.

5.2.3 Sample Training Network

The construction of the BP neural network starts after determining the sample content. The input data were first divided into training and testing sets, the samples were normalized, and the number of iterations, target training error, and learning rate were set. Then start training the network, followed by back normalizing the trained data and sorting them according to the water quality level from 1 to 4.

5.3 Water quality evaluation and analysis

The water quality in the testing area is categorized into four classes from 1 to 4, where class 1 is the best and class 4 is the worst, as specified in the testing index for water quality in industrial water purification plants. The simulation is carried out using the data of a whole year 2022, the data of the first 285 days is used as the training set, and the data of the last 80 days is used as the test set, and the results are shown as follows Fig.5.4-5.5. where the horizontal coordinate represents the data samples used for the assessment, the vertical coordinate represents the water quality classes assessed (1,2,3,4), and the correctness rate is the amount of data correctly assessed for the water quality classes as a percentage of the total amount of data.

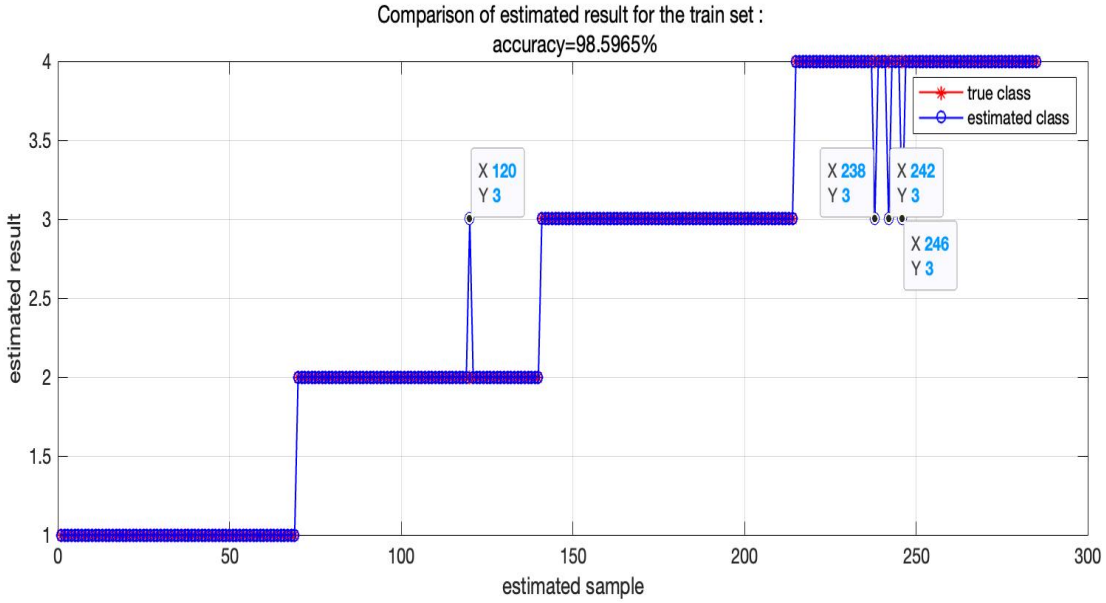

Fig.5.4 Water quality assessment results for the train set

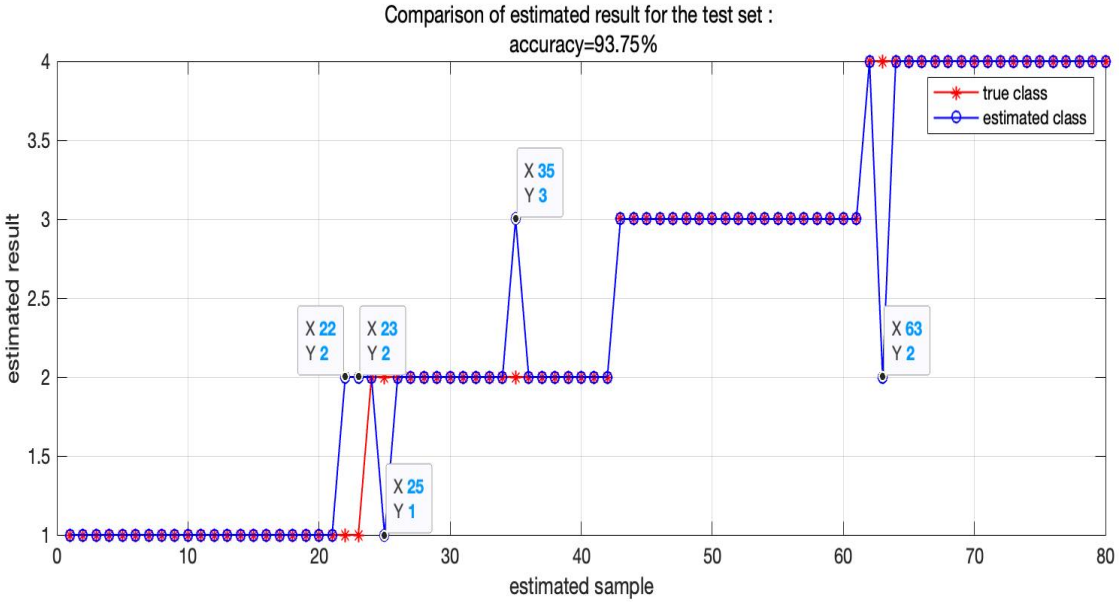

Fig.5.5 Water quality assessment results for the test set

Fig.5.6, 5.7 show the confusion matrices for the training and test sets. It represents the distribution of the evaluation class and the real class. From the figure, it can be seen the amount of data in each class in the evaluation class and the real class and the data that are misclassified.

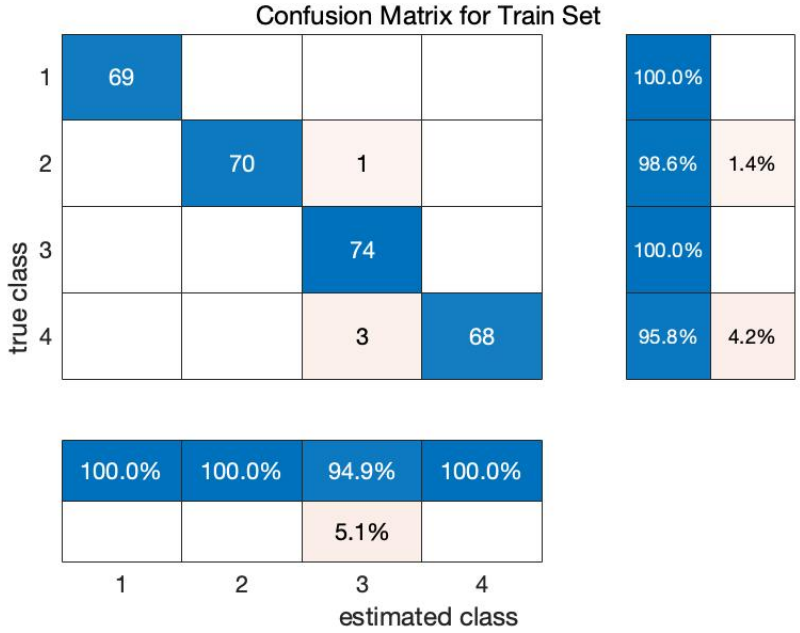

Fig.5.6 Confusion Matrix for the train test

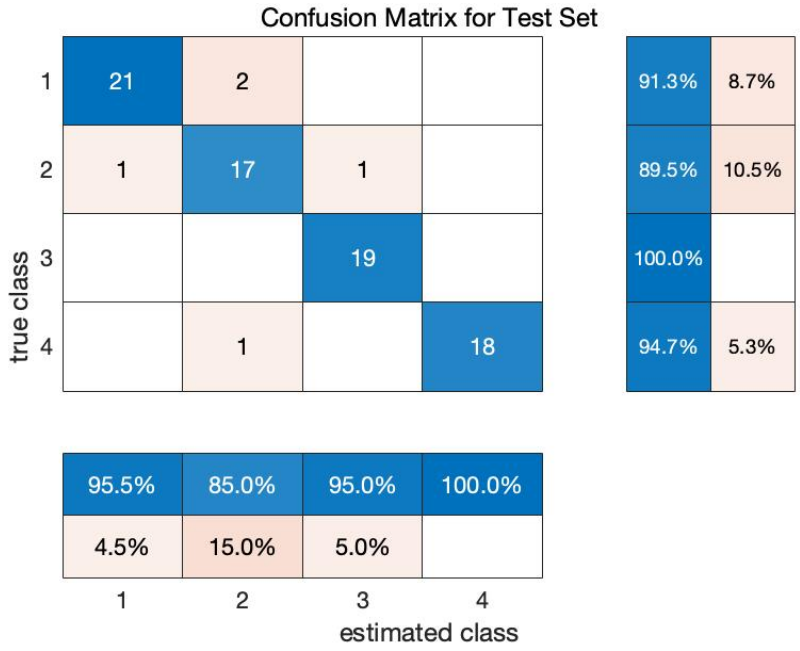

Fig.5.7 Confusion Matrix for the test set

Based on the simulation results, in the training set, the highest false detection rate occurs in data corresponding to water quality level 4, with 3 sets being incorrectly assessed as level 3. In the test set, there are minor false detection rates for water quality levels 1, 2, and 4, but the overall assessment is relatively accurate. This indicates that the feature-level data fusion method based on the BP neural network is effective and feasible for water quality evaluation.

6. Water quality prediction based on LSTM

6.1 LSTM neural network

Long short-term memory (LSTM) is an improved result of traditional recurrent neural networks (RNN). It is a long-term and short-term memory network. Compared with ordinary RNN, LSTM adds a memory cell to judge whether the information is useful or not, which solves the problems of gradient disappearance and gradient explosion in the process of long sequence training[22][23]. This improvement enables it to perform better in longer sequences.

The structure of the LSTM network is shown in Fig.6.1. Its core elements are the cellular state and the gating structure. Cell state is the pathway through which information can be passed down a sequence chain; it can be thought of as the memory of a network.

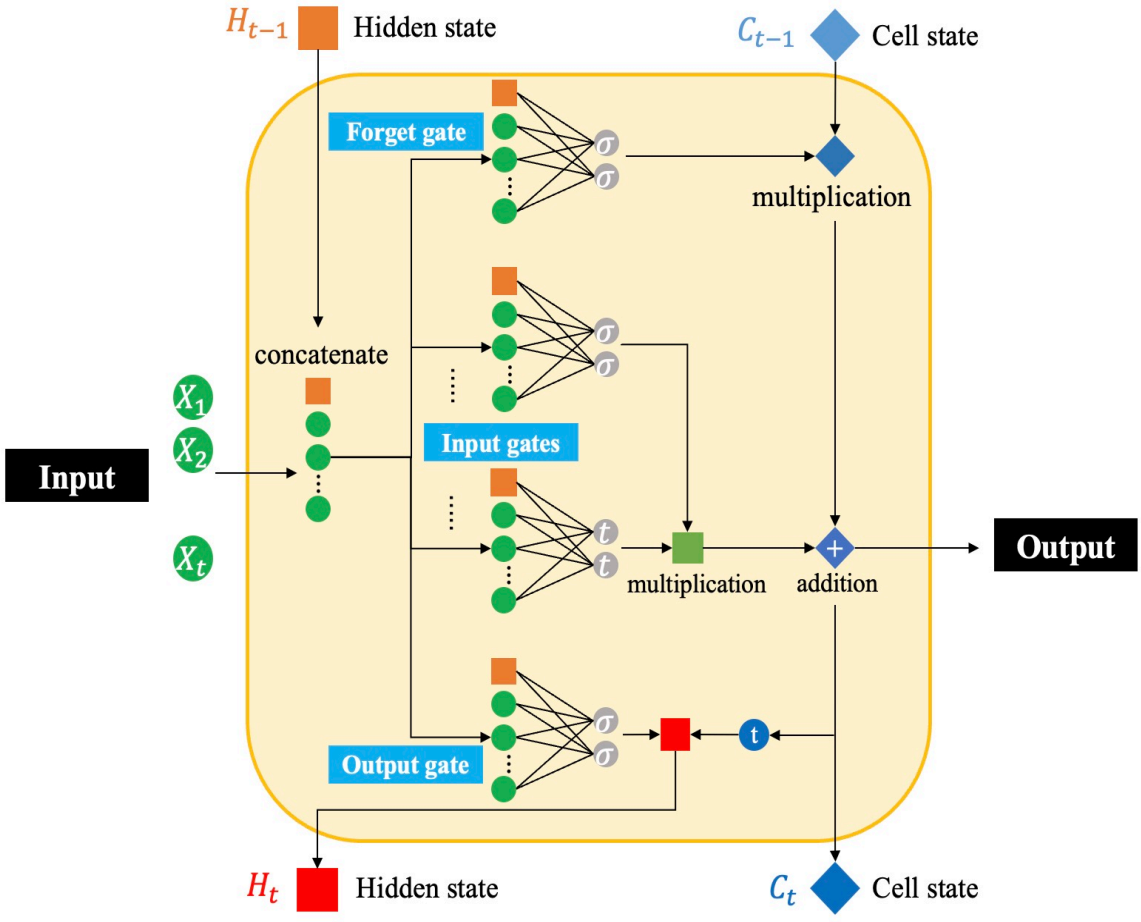

Fig.6.1 LSTM structure

Gating structure generally includes three types of gates: a forget gate, an input gate, and an output gate. Each of these three gates, and the cell state, are described below.

Forget gate (F): Its function is to decide what information should be discarded or retained. The forget gate controls the hidden cell state of the upper layer with a certain probability, and its calculation is shown in Equation 6-1.

$$F_t = \sigma\left(X_t w_f + H_{t-1} w_{fh} + C_t w_{fc} + b_f\right) \quad (6-1)$$

Input gate (I): It is used to update the cell state. The input gate processes the input at the current sequence position and consists of two parts, the results of which are multiplied to update the cell state. Its calculation is shown in Equation 6-2.

$$I_t = \sigma(X_t w_i + H_{t-1} w_{ih} + C_{t-1} w_{ic} + b_i) \quad (6-2)$$

Cell state (C): The cell state depends on the result of the previous forget and input gates, and its calculation is multiplied point-by-point by the cell state of the previous layer and the forget vector, as shown in Equation 6-3.

$$C_t = F_t * C_{t-1} + I_t * \tanh\left(X_t w_c + H_{t-1} w_{ch} + b_c\right) \quad (6-3)$$

Output gate (O): At first, the value of the next hidden state is determined; this contains the information entered earlier. The hidden state is then used as the output of the current cell, and the new cell state and the new hidden state are passed to the next time step. The calculation process is shown in Equations 6-4 and 6-5.

$$O_t = \sigma(X_t w_o + H_{t-1} w_{oh} + C_{t-1} w_{oc} + b_0) \quad (6-4)$$

$$F_t = O_t * \tanh\left(C_{t-1}\right) \quad (6-5)$$

For Equation 6-1 through 6-5, X_t are the input variables; σ represents the sigmoid function; w_f, w_i, w_c, and w_o are the weights of X_t in the forget gate, input gate, cell state, and output gate, respectively; w_{fh}, w_{ih}, w_{ch}, and w_{oh} are the weights of H_{t-1} in the forget gate, input gate, cell state, and output gate, respectively; w_{fc}, w_{ic}, and w_{oc} are weights related to the connection between the cell state and forget gate, input gate, and output gate, respectively; b_f, b_i, b_c, and b_o are the biases in the forget gate, input gate, cell state, and output gate, respectively; and * represents the scalar product of two vectors.

The back propagation algorithm is employed by LSTM in the entire training procedure, and the corresponding parameter matrix will be continuously optimized to finally find a set of optimal parameters.

6.2 Water quality prediction

6.2.1 Single parameter prediction

The simulation data is derived from the water quality monitoring open data of an industrial wastewater treatment plant between 2020 and 2022, comprising a total of 1095 sets. The main parameters include water temperature, pH, turbidity, color, and conductivity.

$$S_{i,n} = \{(y_{i,1}, T_1), (y_{i,2}, T_2), \dots, (y_{i,k}, T_k), \dots, (y_{i,n}, T_n)\} \quad (6-6)$$

Where $y_{i,k}$ is the value of the ith water quality parameter detected by the node at the moment T_k $(1 \le i \le j, 1 \le k \le n)$.$T$ is a time variable, and the sampling interval is fixed as $\Delta T = T_{k+1} - T_k$.for any moment T_k.

For a single parameter $S_{i,n}$ define the forecasting step of the time series as $m, m \in \{1,2, \dots\}$,The single-parameter time series forecasting can be realized by applying the water quality forecasting fusion algorithm, whose task is to give the time series at a future moment:

$$S_{i,n+m} = \{(y_{i,n+1}, T_{n+1}), (y_{i,n+2}, T_{n+2}), \dots, (y_{i,n+k}, T_{n+k}), \dots, (y_{i,n+m}, T_{n+m})\} \quad (6-7)$$

The general processing flow is shown in the Fig.6.2. The processing object of this part is a single environmental parameter in the monitoring sub-area. Take temperature as an example, the first k data collected by the temperature sensor is used as the first temperature time series, and the k+1-th data is predicted through LSTM, and then the second to the k+1-th data predicted just now is used as a new temperature time series, and the k+2-th data is predicted by the same method. The time series is constantly updated by adding the predicted values from the previous time series to the next new time series, this is repeated until k+m data is finally predicted, at which point all data is processed.

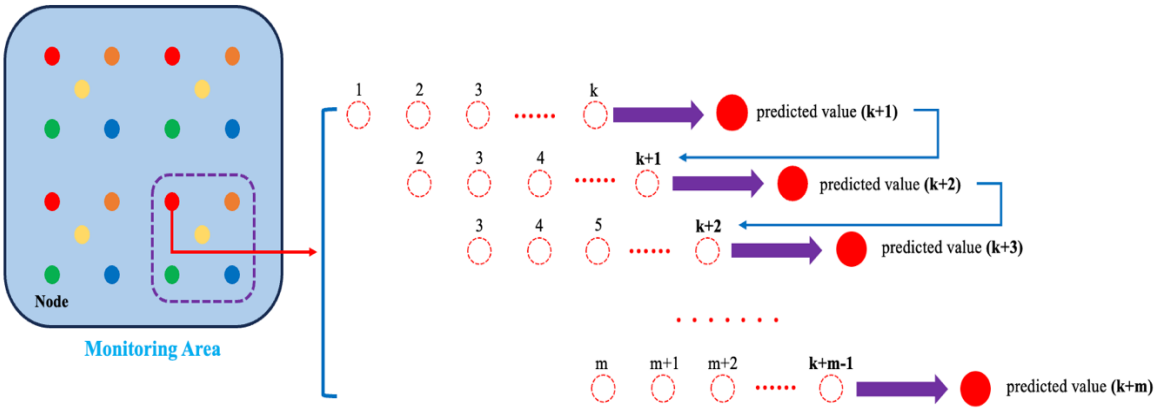

Fig.6.2 Decision level fusion (single parameter) diagram

Based on the analysis, in the fusion of single parameter water quality prediction, five parameters are selected as the fusion targets. A total of 900 sets of data from April 2020 to September 2022 are used as historical data for deep learning and training. The goal is to predict the water quality data for the next 6 months.

(1) Training model parameter settings:

This section sets the training of the prediction model to learn the data patterns for the next 180 days (m=180) based on the historical data of the past 900 days. The training is performed for 100 iterations using the Mini-batch Gradient Descent (MBGD) method, where the training set is divided into mini-batches, and the gradients are computed, and parameters are updated for each batch. The Root Mean Squared Error (RMSE) is used as the loss function, and the Adam optimizer is employed to adjust the LSTM model and update the weights and bias parameters. Once the model training is completed, it can be used to predict the data for each parameter for the following six months starting from September 2022.

(2) Water quality prediction effectiveness：

Using the constructed LSTM-based water quality prediction model, it is possible to achieve predictions of the trend of a single water quality parameter to some extent. However, there may be significant discrepancies in the numerical values compared to the actual data. Fig.6.3, 6.4 show the predicted results for pH and Fig.6.5,6.6 show the predicted results for temperature. The horizontal coordinates in each figure indicate the predicted sample data

and the vertical coordinates indicate the predicted values.

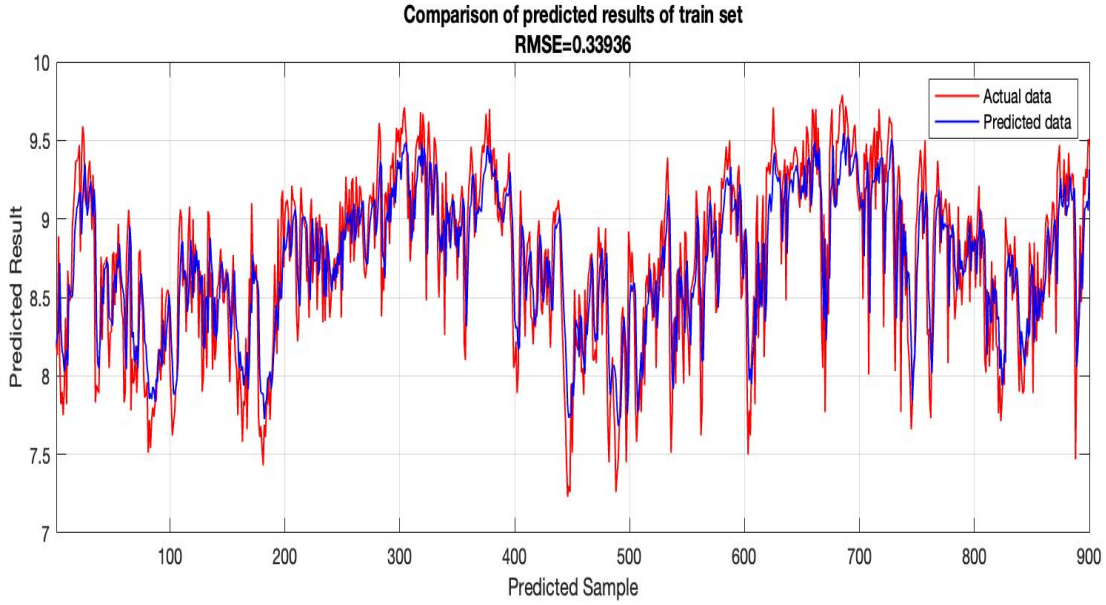

Fig.6.3 Prediction results for the pH single-parameter train set

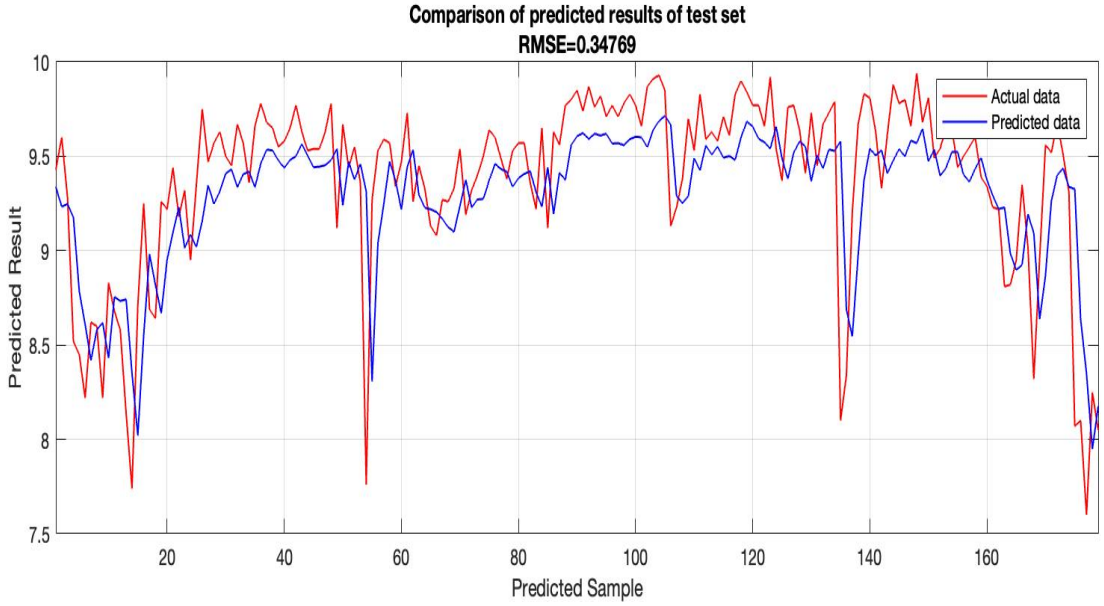

Fig.6.4 Prediction results for the pH single-parameter test set

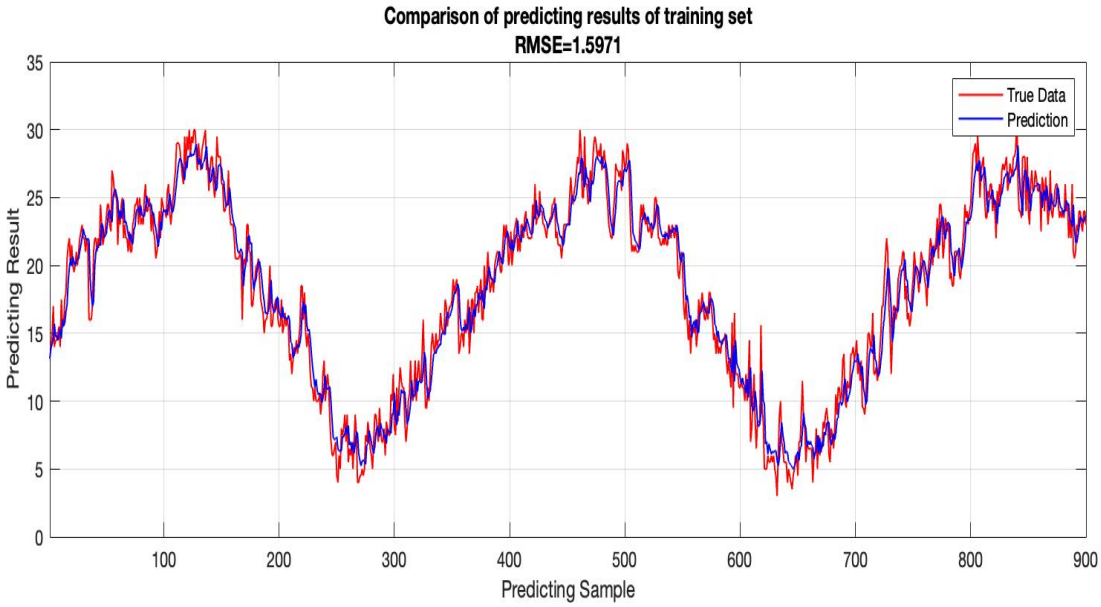

Fig.6.5 Prediction results for the temperature single-parameter train set

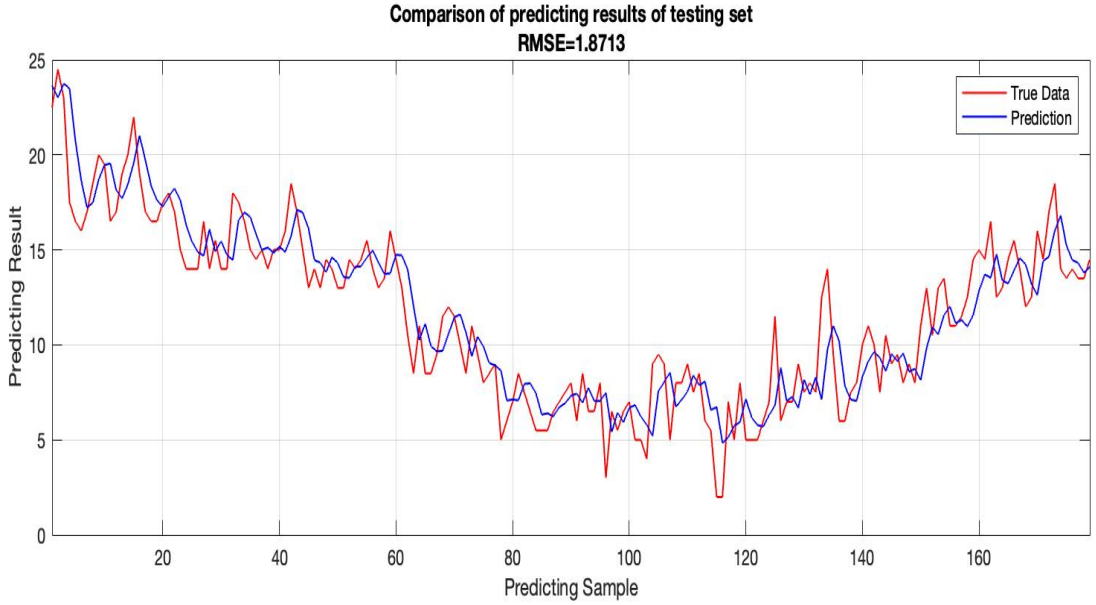

Fig.6.6 Prediction results for the temperature single-parameter test set

6.2.2 Multi parameter prediction

Unlike single parameter, multi parameter time series have more than one time-dependent quantity, i.e., each variable not only depends on its own past values but also has some dependence on other variables, and this dependence between multiple variables can be better

used to predict future values.

For a given water quality multi-parameter time series moment S_n of length n can be defined in the following form:

$$S_n = \begin{bmatrix} (y_{1,1}, T_1) & \cdots & (y_{1,n}, T_n) \\ \vdots & \ddots & \vdots \\ (y_{p,1}, T_1) & \cdots & (y_{p,n}, T_n) \end{bmatrix} \quad (6-8)$$

Define the time prediction step $m, m \in \{1,2, \dots\}$, and p is the number of parameters involved in the water quality prediction ($1 \leq i \leq p$).The main task of using the water quality prediction fusion algorithm for the prediction of a multi-parameter time series is to analyze all the variables related to the predicted parameter to give the future time series of this parameter, which is consistent with a single parameter.

$$S_{i,n+m} = \{(y_{i,n+1}, T_{n+1}), (y_{i,n+2}, T_{n+2}), \dots, (y_{i,n+k}, T_{n+k}), \dots, (y_{i,n+m}, T_{n+m})\} \quad (6-9)$$

The general processing flow is shown in the Fig.6.7. The processing object of this part is all the environmental parameters in the monitoring subarea. All the environmental parameters at the same moment are taken as a group ($r_1 - r_p$, There are p groups), and the group of the first k rounds ($r_1 - r_k$) is taken as the first time series, to predict the value of k+1 data. Then, r_2 to r_{k+1} are used as a second new time series to predict the value of k+2. Each time series contains k groups, the time series is constantly updated by adding the predicted values from the previous time series to the next new time series, this repetition results in the k+m data point, at which point all groups ($r_1 - r_p$) are involved in the fusion processing.

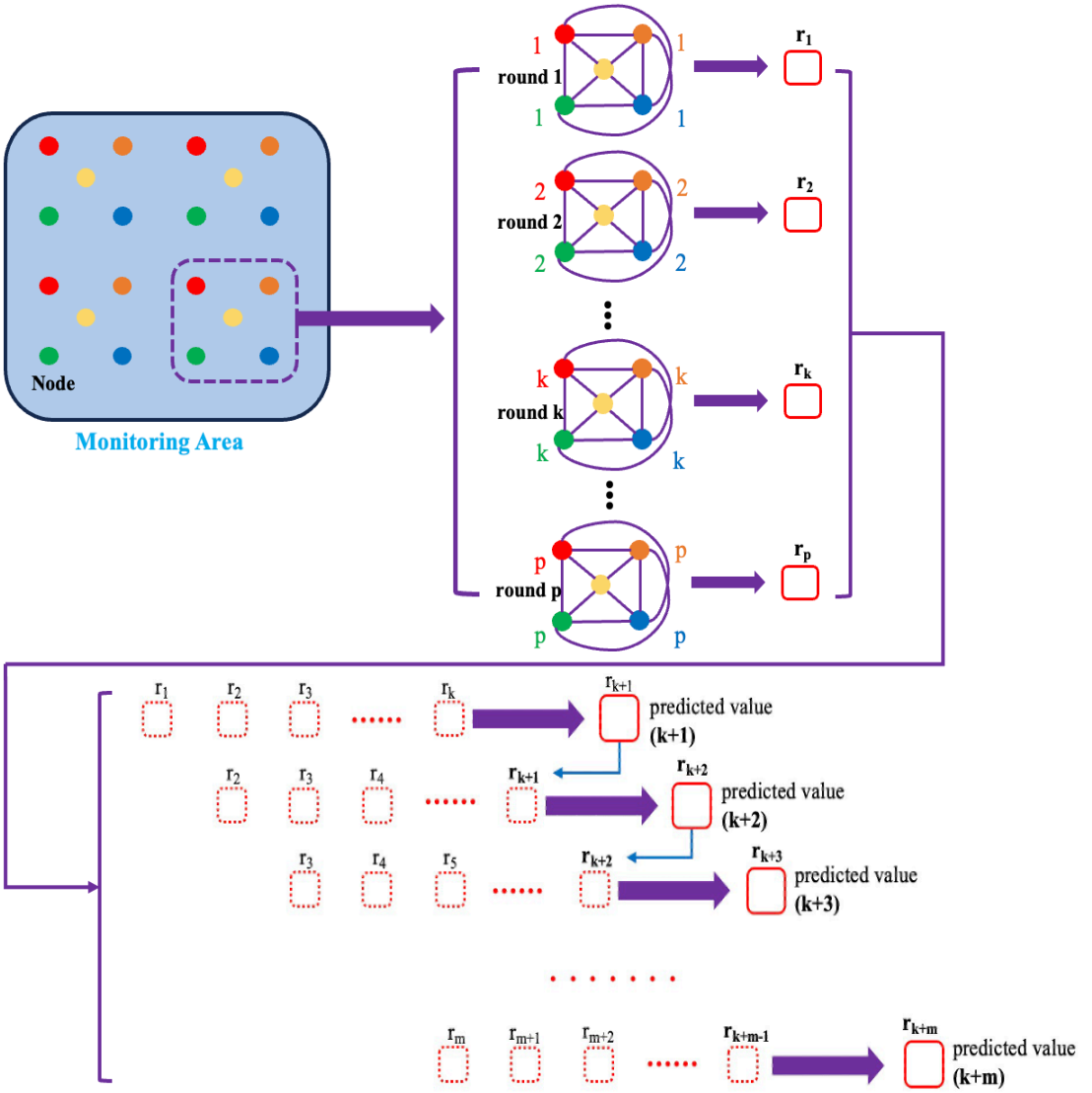

Fig.6.7 Decision level fusion (multi parameter) diagram

The simulation data in this section remains the same as in 6.2.1, although within each parameter's dataset there is historical data for the current parameter in addition to historical data for other parameters with which it may be associated.

(1) Training model parameter settings:

Considering that in multi-parameter prediction, the amount of data involved in learning is significantly increased compared with that of a single parameter, each LSTM layer is set to 128 neurons, batch size is 256, and matrix is 5*15. The RMSE transfer optimization algorithm is used to adjust the model and update the weight and bias parameters, and RMSE is used as the loss function.

(2) Water quality prediction effectiveness：

By utilizing the LSTM-based water quality multi-parameter prediction model, the correlations among multiple parameters can be effectively utilized, leading to a significant improvement in data prediction accuracy. Fig.6.8,6.9 show the prediction results for pH and Fig.6.10,6.11 show the prediction results for temperature. Where the horizontal coordinates indicate the sample data involved in the prediction and the vertical coordinates

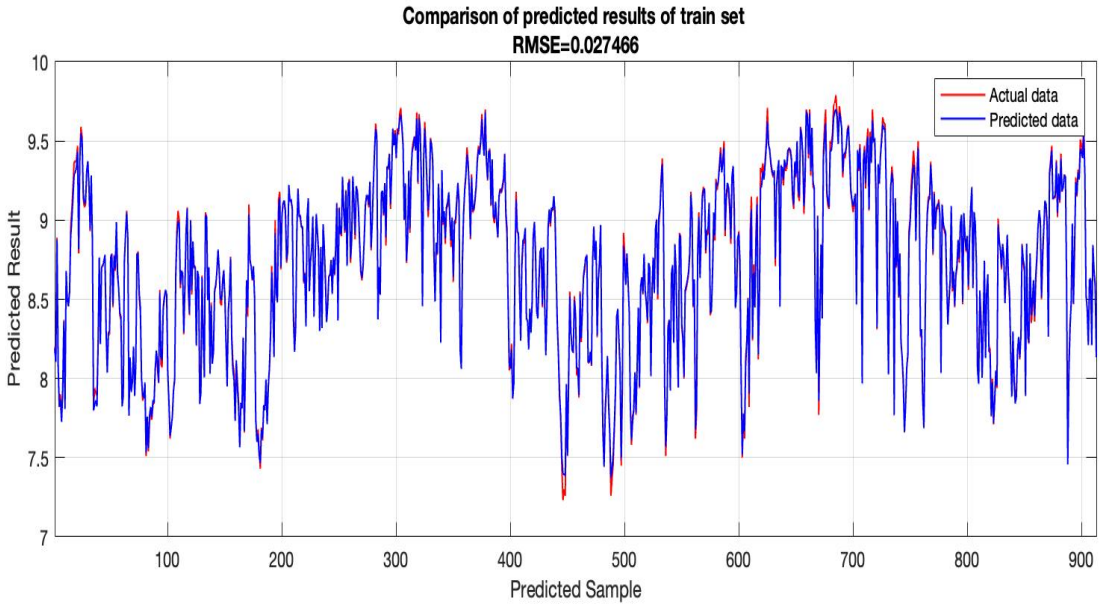

Fig.6.8 Prediction results for the pH multi-parameter train set

Fig.6.9 Prediction results for the pH multi-parameter test set

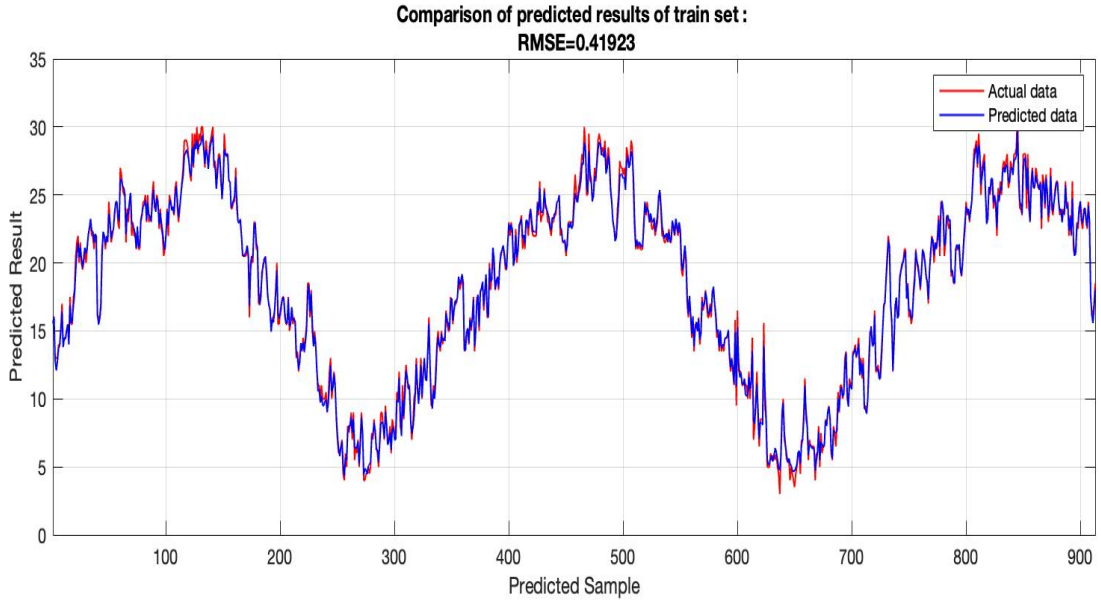

Fig.6.10 Prediction results for the temperature multi-parameter train set

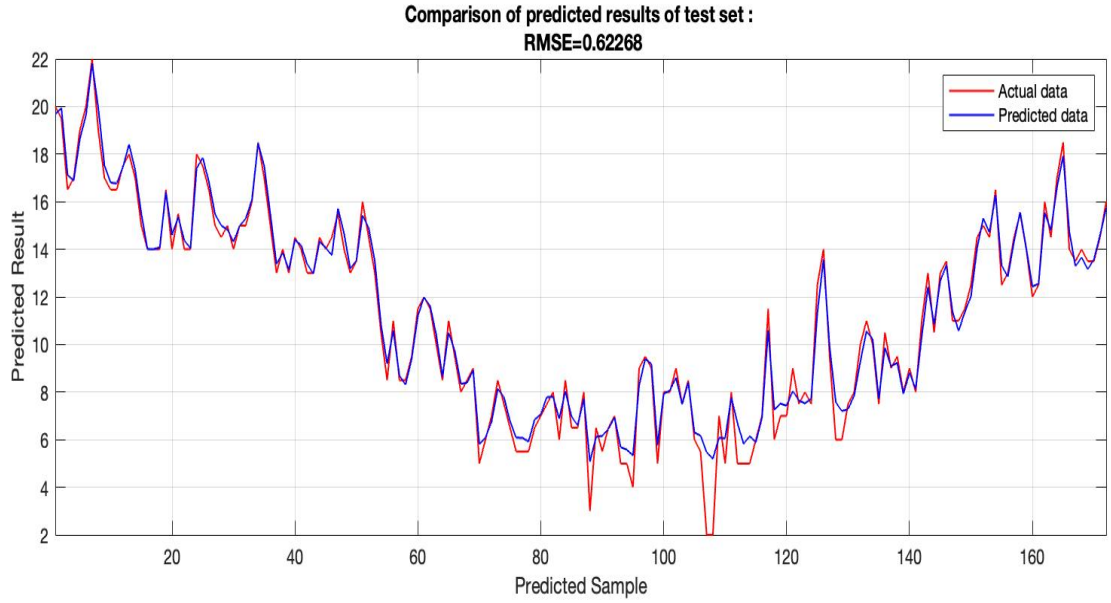

Fig.6.11 Prediction results for the temperature multi-parameter test set

Based on the simulated fusion results, it is evident that the multi-parameter water quality prediction model exhibits good accuracy in estimating water quality over the medium to long term, with minimal deviations from the actual values. This indicates a strong interdependence among the selected parameters within the monitored area.

7. Evaluation

7.1 AWDF data level fusion

The optimized AWDF algorithm proposed in this paper as well as the other two weighted fusion algorithms are simulated and fused separately using MATLAB, data fusion results for temperature are shown in the Fig.7.1. Refer to Table 7.1 for fusion values and total variance.

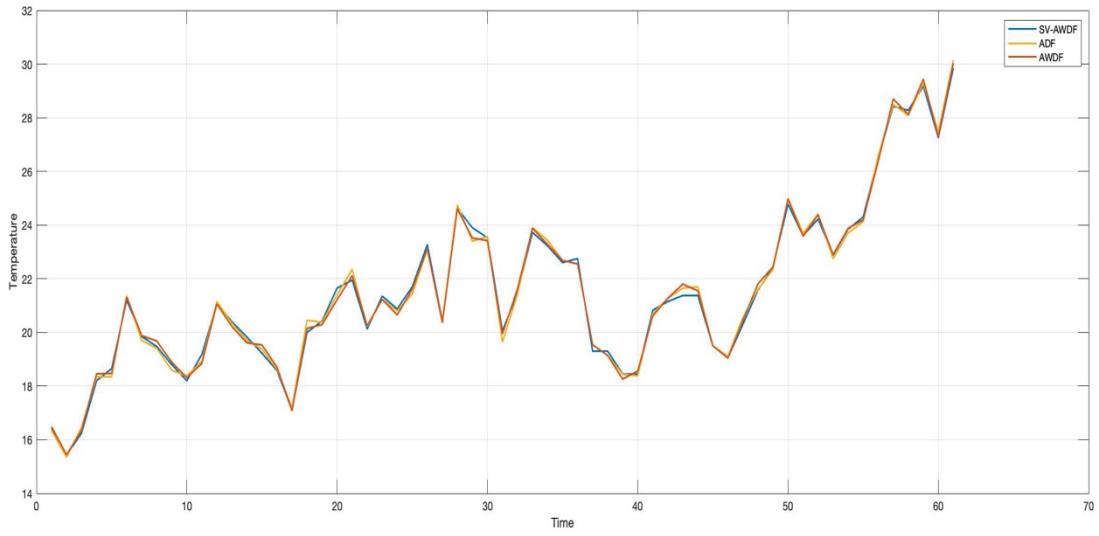

Fig.7.1 Trends in fusion values for the three methods

Table.7.1 Comparison results of three methods

Fusion indicators	ADF	AWDF	SV-AWDF
total variance	9.8324	9.7428	9.6581
fused value	21.7852	21.5186	21.5036

The above comparison shows that although there is not much difference between the proposed SV-AWDF and the other two methods in terms of fusion values, the introduction of virtual sensors makes it possible to satisfy the requirement of minimizing the total variance with a small number of acquisition nodes, which improves the fusion accuracy.

7.2 Neural network Feature level fusion

The simulation results of water quality assessment using a general neural network are shown below:

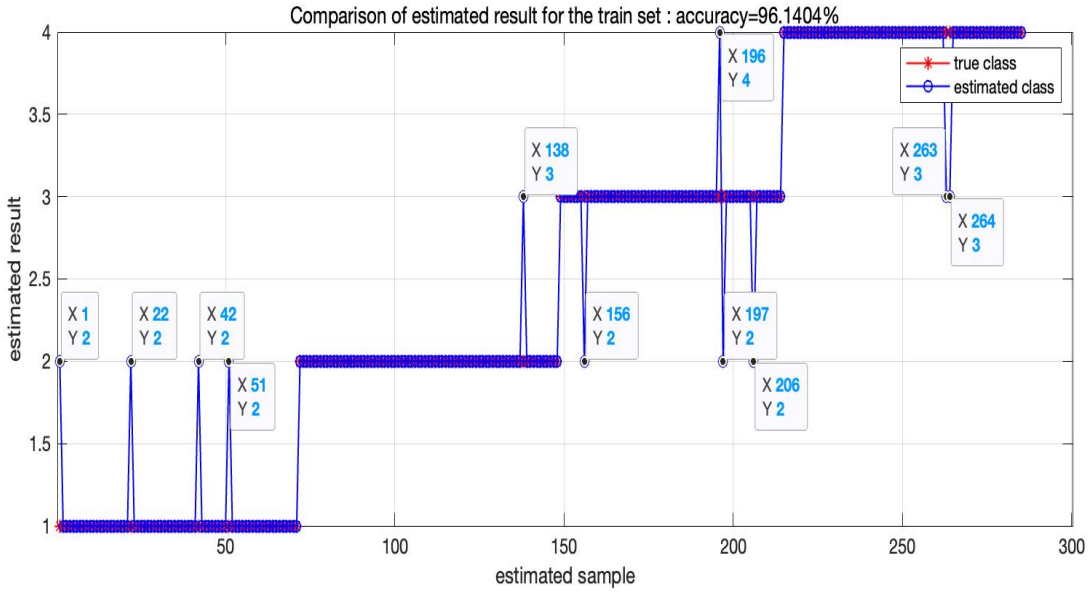

Fig.7.2 Water quality assessment results for the train set

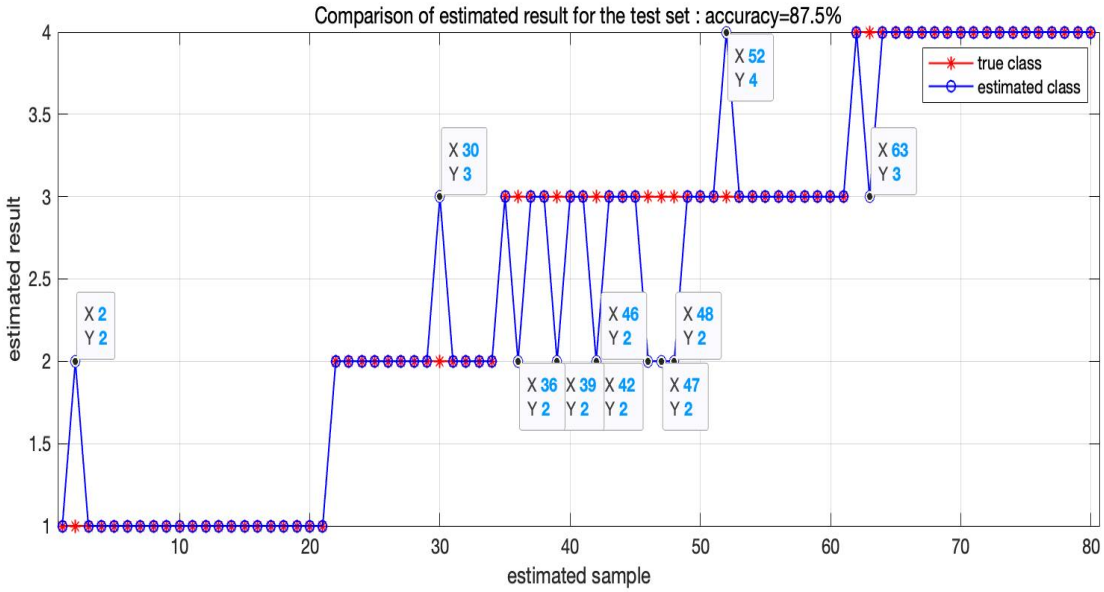

Fig.7.3 Water quality assessment results for the test set

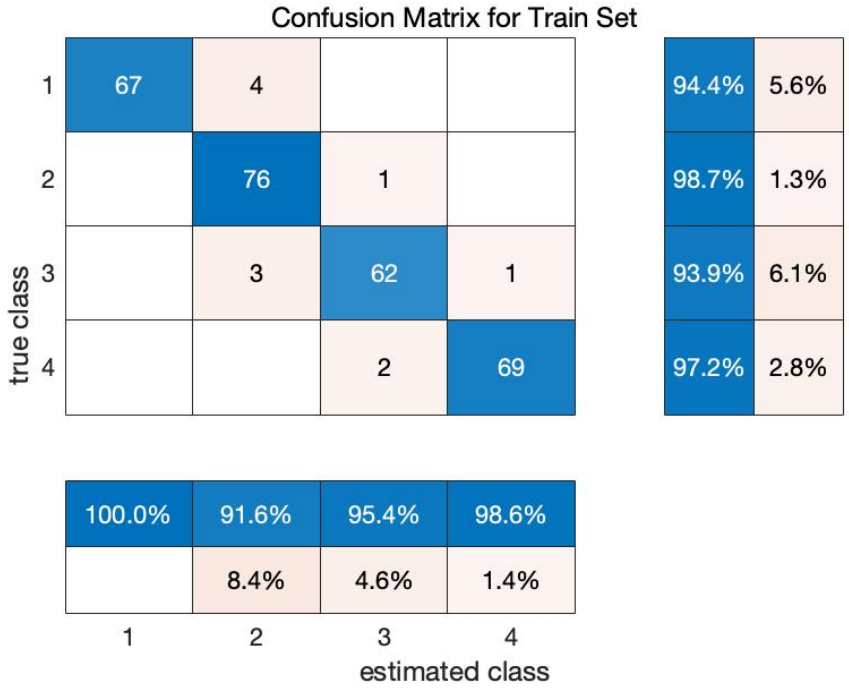

Fig.7.4 Confusion Matrix for the train test

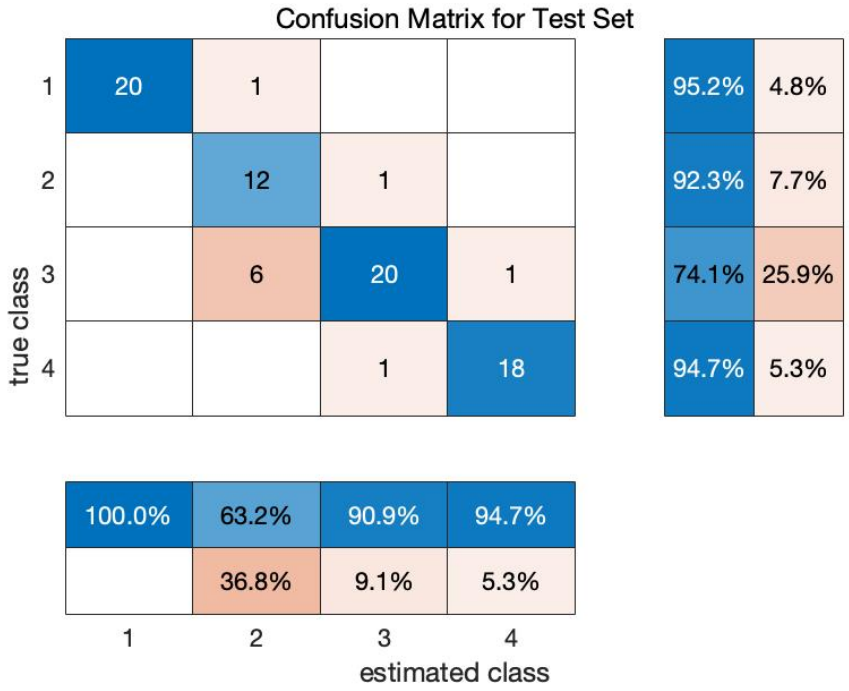

Fig.7.5 Confusion Matrix for the test set

Comparing the above figure, it can be found that the classification of water quality assessment using general artificial neural network is also effective, but the accuracy of

classification is not enough compared to BP neural network. There will be a concentration of misclassification, for example, in Fig.7.3, six data with a true water quality class of 3 were incorrectly assessed as class 2.

7.3 LSTM Decision level fusion

In the final decision level fusion, comparing the single-parameter and multi-parameter water quality prediction results it is easy to find that.

(1) the single-parameter prediction is less satisfactory and has a large error with the actual data. Initial speculation may be that the waters where the data are collected have more obvious changes of their own, and only considering the changes of a single parameter is not enough to predict the future values.

(2) Compared with the instability and large error of single-parameter prediction, multi-parameter prediction can more effectively utilize the possible connection between various water quality indicators to optimize the prediction to the greatest extent.

In addition, to predict water quality parameters more accurately and compare the impact of the amount of data involved in the training on the prediction results, this part adds the water quality data of 2018 and 2019 based on the data from 2020 to 2022, that is, the data of 3 years has been changed to 5 years. As before, it is divided into single parameter and multi-parameter predictions. The difference is that the ratio of the training set to the test set is different, in the 5 years of data, the first 4 years are input as the training set and the last 1 year is output as the test set.

Fig.7.6,7.7 below show the results of the train set and the test set for single-parameter prediction of water temperature respectively. The analysis shows that after adding 2 years of water quality data, the result of using LSTM single parameter to predict water temperature is like that before, and the accuracy is not significantly improved.

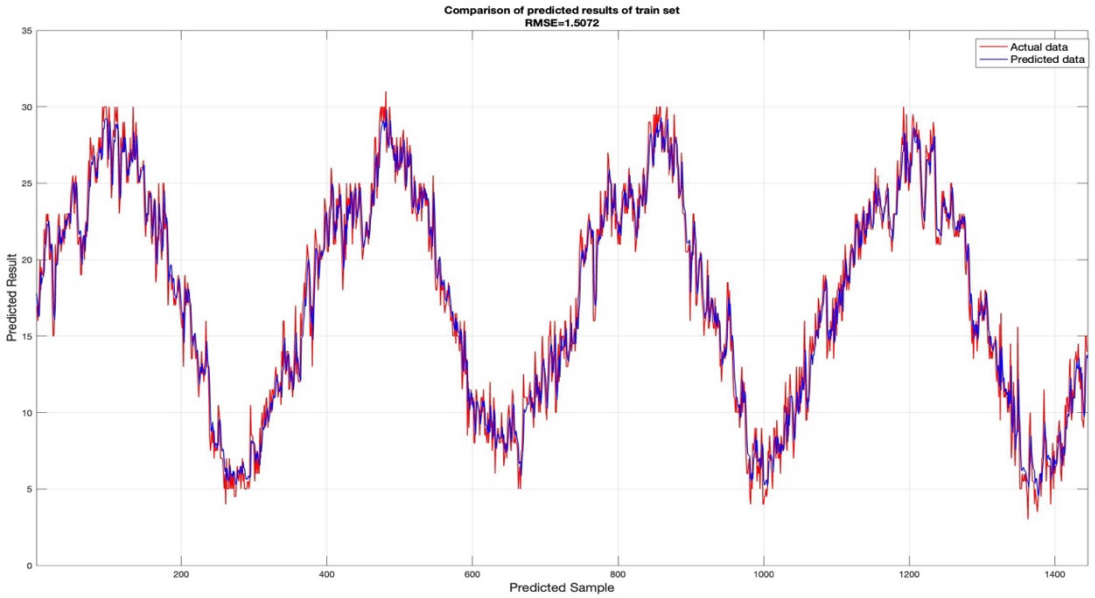

Fig.7.6 Prediction results for the temperature single-parameter train set (2018-2022)

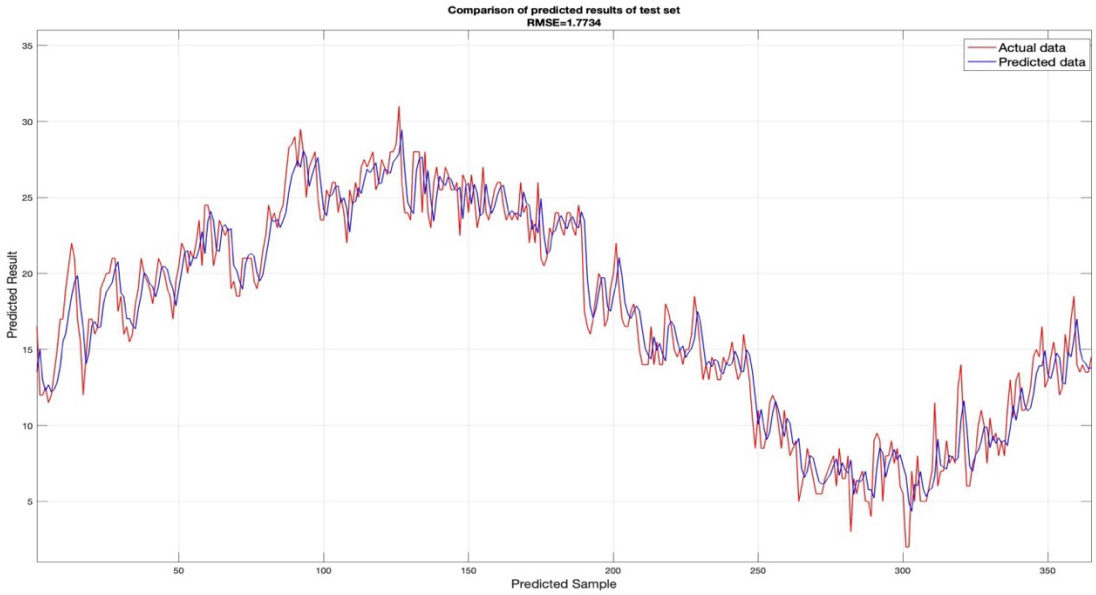

Fig.7.7 Prediction results for the temperature single-parameter test set (2018-2022)

Fig.7.8,7.9 below show the results of the train set and test set for single-parameter prediction of pH, respectively. Like the prediction results for water temperature, after 5 years of water quality data, the prediction results are still like the previous ones, and the accuracy is not significantly improved.

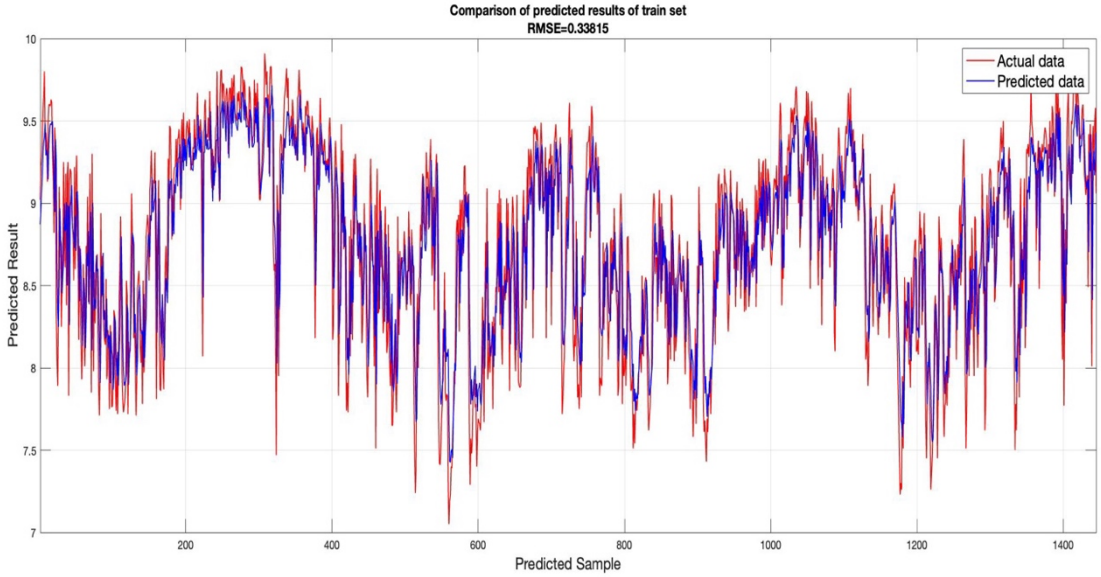

Fig.7.8 Prediction results for the pH single-parameter train set (2018-2022)

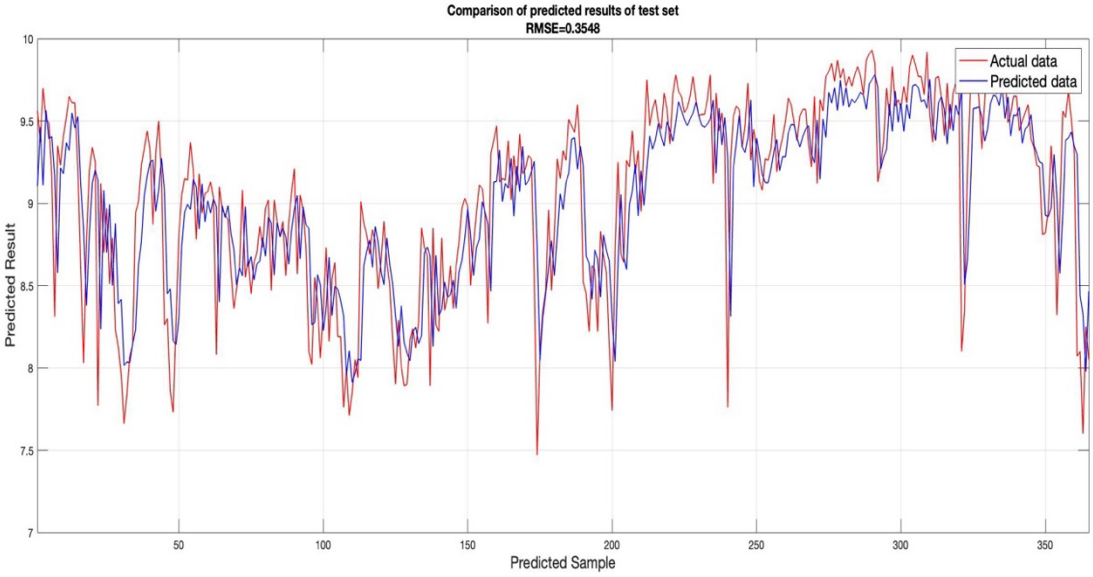

Fig.7.9 Prediction results for the pH single-parameter test set (2018-2022)

Different from single parameter prediction, the accuracy of multi-parameter prediction has been significantly improved after two more years of data. Fig.7.10,7.11 show the results of train set and test set for multi-parameter prediction of water temperature respectively. Fig.7.12,7.13 show the results of the train set and test set for multi-parameter prediction of pH.

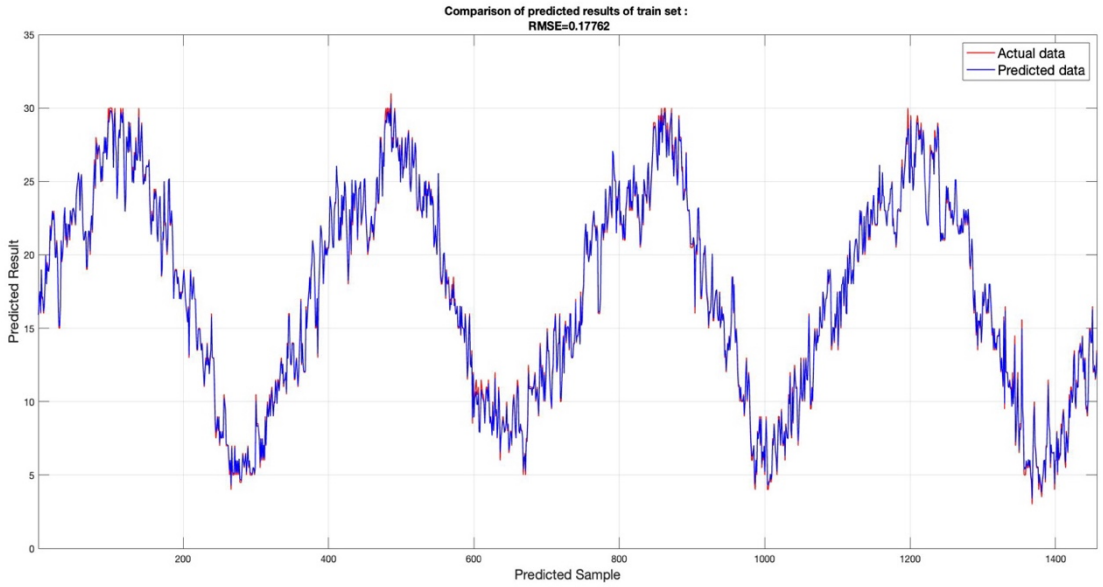

Fig.7.10 Prediction results for the temperature multi-parameter train set (2018-2022)

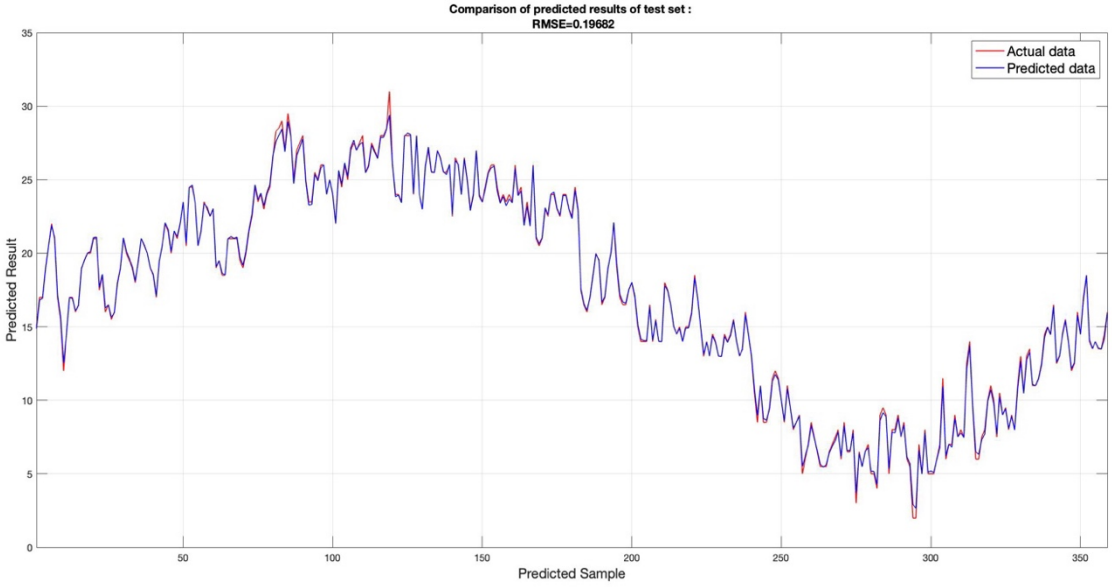

Fig.7.11 Prediction results for the temperature multi-parameter test set (2018-2022)

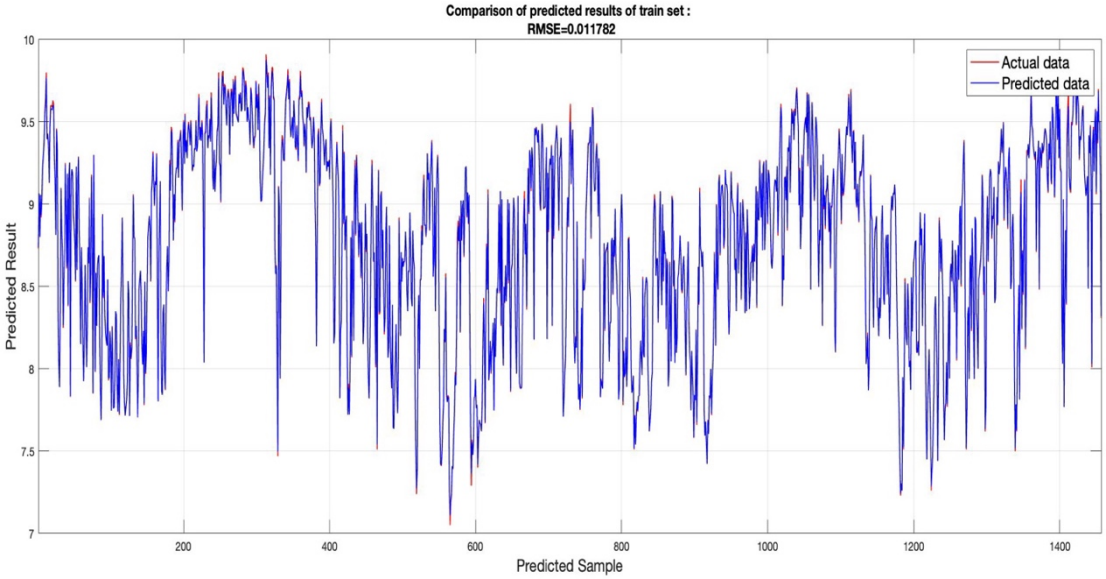

Fig.7.12 Prediction results for the pH multi-parameter train set (2018-2022)

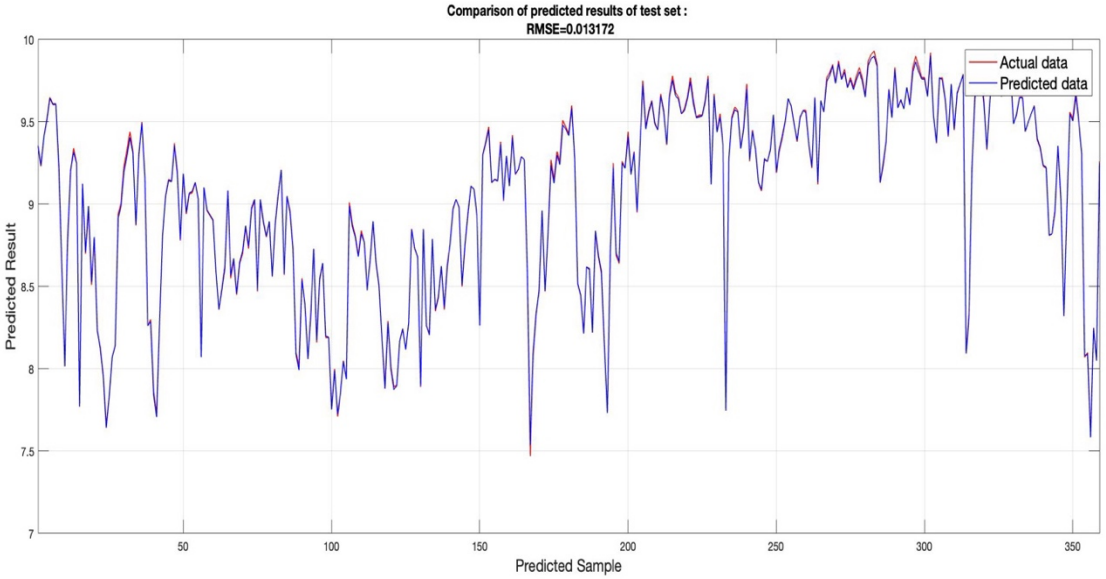

Fig.7.13 Prediction results for the pH multi-parameter test set (2018-2022)

Combining the single and multi-parameter prediction using 2020-2022 water quality data in Section 6.2 with the single and multi-parameter prediction using 2018-2022 water quality data in this section, the RMSE under different conditions is analyzed with the predicted results of water temperature and pH as examples. The following Table 7.2 shows the specific comparison results.

Table 7.2 RMSE of single and multi-parameter prediction for different data quantities

method	RMSE of water quality parameter prediction							
	temperature				pH			
	train set		test set		train set		test set	
	2020-2022	2018-2022	2020-2022	2018-2022	2020-2022	2018-2022	2020-2022	2018-2022
single-parameter	1.5971	1.5072	1.8713	1.7723	0.33936	0.3382	0.34769	0.3548
multi-parameter	0.41923	0.17762	0.6227	0.19682	0.02747	0.0118	0.03814	0.0132

Through the above analysis and comparison, it is not difficult to find that in the aspect of single-parameter prediction of water quality, increasing the amount of data does not significantly improve the accuracy of prediction. However, when using multi-parameter prediction, the increase of data volume significantly improves the accuracy of the final prediction. It can be further learned that when there is some correlation between multiple data involved in training, increasing the amount of data can improve the accuracy of prediction.

8. Conclusion

Water quality monitoring is a basic measure to protect the water environment, through the collection and analysis of various environmental parameters in the target waters, observing the changes in the water environment, providing an effective means for water quality analysis and assessment, water quality prediction and other aspects.

Which is based on WSNs water environment monitoring system has attracted much attention in recent years, with a variety of sensors distributed in the detection of water to form a wireless sensor network, can be more timely and effective monitoring of the target waters. However, if the raw data collected by the sensors are not processed but directly transmitted, then on the one hand, for the entire network, it will increase the amount of transmission data and energy consumption of the nodes, resulting in network congestion; on the other hand, it is difficult to extract useful data from a large number of environmental parameters to analyze the current water environment, which is prone to one-sided assessment of water quality, and will also have an impact on the subsequent prediction of water quality. Therefore, it is necessary to introduce some reasonable data fusion mechanisms to solve these problems. In this article, we consider the tasks faced in the three levels of numerical fusion, feature fusion and decision fusion, corresponding to the adoption of different data fusion methods to improve the problems in water environment monitoring. The specific research results are as follows:

(1) Several topics that need to be dealt with in water environment monitoring based on WSNs are first analyzed, and a new multilevel data fusion method applied to water environment monitoring is proposed.

(2) At the data level, the AWDF algorithm is used to perform a preliminary fusion process on the raw data to reduce the amount of data that needs to be transmitted. However, considering the situation that many sensors may not be able to be placed in the actual monitoring environment, which may lead to low fusion accuracy, this article proposes an SV-AWDF method to improve the number of fusions by adding virtual sensors, which in

turn ensures the accuracy of the results. For the fusion processing of water temperature, pH and other parameters in the water environment, comparing the fusion results of ADF, ordinary AWDF and the SV-AWDF of this article on the original data, the analysis concludes that the improved optimization algorithm improves the accuracy compared with the other two methods, and can be applied to a certain extent to the situation of insufficient number of sensors.

(3) At the characterization level, a neural network-based data fusion method is proposed for water quality parameters (water temperature, pH, turbidity, chromaticity, and conductivity) that may be correlated in the water environment. The aim is to utilize the correlation between these parameters to give a reasonable assessment of water quality. Using 285 days of water quality data as a training sample to assess the last 80 days of water quality, the results show that the accuracy can be as high as 93.75%.

(4) At the decision-making level, to further extend the function of data fusion, this article utilizes LSTM deep neural network to make predictions on medium- and long-term water environment parameters and proposes different data fusion methods for single parameter and multi-parameter respectively. The results show that the multi-parameter can utilize the correlation between various water quality data under certain conditions and can give more accurate predictions than the single-parameter prediction, and the RMSE can be reduced to about 0.027. In the assessment process, water quality data of 2018 and 2019 were added as a comparison. It was found that for single-parameter prediction, increasing the amount of training data could not significantly improve the accuracy of prediction. However, it is different for multiple parameters, and the more samples involved in training, the higher the accuracy of the final prediction.

Prospect:

In this article, the water environment monitoring of WSNs is taken as the research background, and different data fusion methods are proposed in the three aspects of numerical value, feature correlation among data, and data prediction, respectively, considering the water temperature, pH, turbidity, and other parameters of the actual water environment.

Although the simulation results show good performance and provide a reliable method for data analysis of WSNs water environment monitoring system to a certain extent, the following problems still exist:

(1) Due to the limitation of not conducting experiments in the actual water environment, the acquired environmental parameters are not obtained through sensor acquisition, so there is an ideal situation in the collection of raw data (lack of anomalies, missing data, etc.), and it is not possible to further analyze and compare the energy consumption of the sensor network, transmission delay, and so on.

(2) In the prediction of a single water quality parameter, its prediction results differed greatly from the actual data, but the multi-parameter prediction was effective. In addition to the correlation between parameters has a large impact on the prediction results, this paper does not consider the interference of other factors.

Based on the above two shortcomings, after that, when conditions allow, experiments in the actual water environment can more accurately reflect the changes of water quality parameters (considering the sensor's own influence, external interference). In addition, analyzing the excessive error and lack of precision that exists in the prediction of a single parameter is also one of the focuses of future research.

Reference

[1] Trasviña-Moreno C A, Blasco R, Marco Á, et al. Unmanned aerial vehicle based. wireless sensor network for marine-coastal environment monitoring[J]. Sensors, 2017, 17(3): 460.

[2] Jing W, Tingting L. Application of wireless sensor network in Yangtze River basin. water environment monitoring[C]//The 27th Chinese control and decision conference (2015 CCDC). IEEE, 2015: 5981-5985.

[3] Xiao X, Huang H, Wang W. Underwater wireless sensor networks: An energy-efficient. clustering routing protocol based on data fusion and genetic algorithms[J]. Applied Sciences, 2020, 11(1): 312.

[4] Donta P K, Amgoth T, Annavarapu C S R. Delay-aware data fusion in duty-cycled. wireless sensor networks: A Q-learning approach[J]. Sustainable Computing: Informatics and Systems, 2022, 33: 100642.

[5] Zhang Z, Glaser S D, Bales R C, et al. Technical report: The design and evaluation of a basin‐scale wireless sensor network for mountain hydrology[J]. Water Resources Research, 2017, 53(5): 4487-4498.

[6] Zhang X N, Zhao Y, Guo L. Aquatic ecological zoning of SongHua River Basin based on data fusion technology [J]. JOURNAL OF HARBIN INSTITUTE OF TECHNOLOGY, 2019, 51(8): 80-87.(in Chinese)

[7] Raza U, Camerra A, Murphy A L, et al. Practical data prediction for real-world wireless sensor networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(8): 2231-2244.

[8] Horita F E A, de Albuquerque J P, Degrossi L C, et al. Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with wireless sensor networks[J]. Computers & Geosciences, 2015, 80: 84-94.

[9] Furquim G, Pessin G, Faiçal B S, et al. Improving the accuracy of a flood forecasting

model by means of machine learning and chaos theory: a case study involving a real wireless sensor network deployment in brazil[J]. Neural computing and applications, 2016, 27: 1129-1141.

[10] Luo L, Zhang Y, Zhu W. E-Science application of wireless sensor networks in eco-hydrological monitoring in the Heihe River basin, China[J]. IET Science, Measurement & Technology, 2012, 6(6): 432-439.

[11] Sun G, Zhang Z, Zheng B, et al. Multi-sensor data fusion algorithm based on trust degree. and improved genetics[J]. Sensors, 2019, 19(9): 2139.

[12] Gong L, Yan J, Chen Y, et al. An IoT-based intelligent irrigation system with data fusion. and a self-powered wide-area network[J]. Journal of Industrial Information Integration, 2022, 29: 100367.

[13] Khudonogova L I, Muravyov S V. Interval data fusion with preference aggregation for balancing measurement accuracy and energy consumption in WSN[J]. Wireless Personal Communications, 2021, 118: 2399-2421.

[14] Ullah I, Youn J, Han Y H. Multisensor data fusion based on modified belief entropy in. Dempster–Shafer theory for smart environment[J]. IEEE Access, 2021, 9: 37813-37822.

[15] Yang M R，Guo X F，Huang Y F，et al.WSN water quality anomaly detection. technology based on data compression[J]. Video Engineering，2022，46（5）：204-207.

[16] Zhu X N. Research on Data Fusion Algorithm Based on Neural Network in. Wireless Sensor Network[D]. Jilin University,2016. (in Chinese)

[17] Li D, Shen C, Dai X, et al. Research on data fusion of adaptive weighted multi-source ensor[J]. Computers, Materials & Continua, 2019, 61(3): 1217-1231.

[18] Zhao Y, Yang X, Wu X, et al. Adaptive Weighting Strategy based Multi-sensor Data. Fusion Method for Condition Monitoring of Reciprocating Pump[C]//2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS). IEEE, 2021: 1-6.

[19] Chiba prefecture open data site, 「Water quality information」 https://www.pref.chiba.lg.jp/kigyou/kyshisetsu/opendata/opendata-suishitsu.html

[20] Abiodun O I, Jantan A, Omolara A E, et al. State-of-the-art in artificial neural network. applications: A survey[J]. Heliyon, 2018, 4(11).

[21] Saritas M M, Yasar A. Performance analysis of ANN and Naive Bayes classification. algorithm for data classification[J]. International journal of intelligent systems and applications in engineering, 2019, 7(2): 88-91.

[22] Yin H, Li D, Wang Y, et al. Adaptive Data Fusion Method of Multisensors Based on LSTM-GWFA Hybrid Model for Tracking Dynamic Targets[J]. Sensors, 2022, 22(15): 5800.

[23] Staudemeyer R C, Morris E R. Understanding LSTM--a tutorial into long short-term. memory recurrent neural networks[J]. arXiv preprint arXiv:1909.09586, 2019.

Acknowledgments

Time passes quickly like a white pony 's shadow across a crevice, and in the summer of 2023, my campus life belonging to me finally came to an end. Looking back on these two and a half years, I can say that most of time were spent in the COVID-19. There are too many happy and unforgettable good times to remember, and at the same time there are many painful and sad difficult moments that haunt my heart. I would like to thank all of you who have helped me. Your encouragement and support have helped me to sustain myself until today.

During my two and a half years as a graduate student, I am grateful to Prof.Ohshima for giving me this opportunity, and his continuous care and help to my life and study.

Thank my good friend Jiang Yihao, you helped me a lot in my research and gave me useful tips for reading article.

Thank my roommate Zhuang Yu, you taught me many valuable life lessons even though I only spent half a year with you. Although you were younger than me, you took care of me and encouraged me in every way.

Finally, I would like to thank my parents, who have been silently encouraging and helping me on my way to study, which is the driving force for me to keep moving forward.

The time spent in Japan will be the most unforgettable memory in my life. I will take everything I have gained in Japan and set out for a new future.