Hindawi
Journal of Spectroscopy
Volume 2021, Article ID 5957376, 17 pages
https://doi.org/10.1155/2021/5957376

Research Article

Evaluation of the Effectiveness of Multiple Machine Learning Methods in Remote Sensing Quantitative Retrieval of Suspended Matter Concentrations: A Case Study of Nansi Lake in North China

Xiuyu Liu,[1] **Zhen Zhang**,[1] **Tao Jiang**,[1] **Xuehua Li**,[1] **and Yanyi Li**[2]

[1]*College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266500, China*
[2]*College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China*

Correspondence should be addressed to Tao Jiang; tj801020921@sdust.edu.cn

Received 20 April 2021; Revised 30 July 2021; Accepted 5 August 2021; Published 14 August 2021

Academic Editor: Arnaud Cuisset

Total suspended matter (TSM) is a core parameter in the quantitative retrieval of ocean color remote sensing and an important indicator for evaluating the quality of the aquatic environment. This study selects part of Nansi Lake in North China as the study area. Researchers used Hyperion remote sensing data and field-measured TSM concentration as data sources. Firstly, the characteristic variables with high correlation were selected based on spectral analysis. Then, seven methods such as linear regression, BP neural network (BP), KNN, random forest (RF), and random forest based on genetic algorithm optimization (GA_RF) are used to construct the inversion model of TSM concentration. The retrieval accuracy of each model shows that the machine learning models are much more accurate than the linear model. Among them, the GA_RF model retrieves the suspended solids concentration with the best performance and the highest prediction accuracy, with a determination coefficient R^2 of 0.98, a root mean square error (RMSE) of 1.715 mg/L, and an average relative error (ARE) of 6.83%. Additionally, the spatial distribution of TSM concentration was inversed by Hyperion remote sensing image. The results showed that the concentration of TSM was lower in the northwest and higher in the southeast, and the concentration distribution was uneven, showing the characteristics of a typical shallow macrophytic lake. This study provides an effective method for monitoring TSM concentration and other water quality parameters in the shallow macrophytic lake and further proves the advantages of machine learning in ocean color inversion. All in all, this research provides some useful methods and suggestions for quantitative inversion of TSM concentration in shallow macrophytic lakes.

1. Introduction

TSM is one of the important parameters of water quality, which directly affects the transparency and turbidity of the water body [1] and then affects the growth of aquatic organisms and the primary productivity of the water body [2]. Traditional water quality monitoring mainly adopts the method of fixed section and fixed point for sampling analysis, which not only is costly and greatly affected by external conditions but also has defects in large-scale real-time detection. As one of the important methods of environmental investigation and detection, remote sensing technology can quickly obtain water quality information of large areas of water. With the advantages of rapid, real-time, large-scale, periodic dynamic detection, remote sensing has become an important method to detect the spatial and temporal distribution of TSM [3].

Ocean color remote sensing provides an effective way to monitor the three components of ocean color and their spatiotemporal changes. The sensor receives the spectral signal from the water body, analyzes the element information in the water, and then establishes a model to retrieve the component concentration in the water. The element information refers to the concentration of certain substances in the water, such as the concentration of suspended matter, the concentration of chlorophyll, and the concentration of CDOM (Colored Dissolved Organic Matter). The remote sensing retrieval methods of TSM concentration in water

bodies can be divided into analytical, semianalytical, and empirical methods. The core of the above analysis method is the bio-optical model [4], which is based on the radiative transfer theory. The theory mainly analyzes the absorption characteristics of upward radiation and optically active substances in water and analyzes the relationship between these characteristics and backscattering characteristics. Furthermore, the remote sensing reflectance spectrum characteristics of the actual water body are combined with the inherent optical quantity of water components, and then the water quality parameters are retrieved [5]. This type of method has a clear physical meaning and the retrieval results are more reliable, but it requires measurements of the intrinsic optical properties of water components and the algorithms are difficult to establish. The semianalytical method is a combination of known spectral characteristics of water quality parameters and statistical models to select the optimal band or a combination of bands as variables for remote sensing retrieval, which has strong reliability and applicability and is widely used in the estimation of parameters of Case II water [6, 7]. The empirical method is based on the statistical relationship between measured TSM concentration data and remotely sensed spectral information to retrieve TSM concentrations. It is the most commonly used ocean color remote sensing retrieval algorithm at present due to its relatively simple method and fewer parameters. The empirical methods search for the best band or band combination to build a mathematical model, mainly including single-band model [8, 9], band ratio model [10], and multiple regression model [11]. With the continuous development and combination of computer technology and remote sensing technology, the application of machine learning in the field of remote sensing image processing has gradually become a research hotspot. Case II waters, such as rivers, lakes, and near-shore marine waters, are optically complex and poorly suited to the use of linear models for studying Case I waters in isolation [12]. Machine learning has the advantage of solving complex nonlinear problems. For example, Louis et al. combined a neural network model with TM images to simulate suspended sediment concentrations, and the results demonstrated that the network model performs better on nonlinear problems compared to traditional regression analysis [13]. Chen et al. used a multilayer backpropagation neural network (MBPNN) model to estimate TSS concentrations in the coastal zone of eastern China, and obtained results were all more accurate than those of empirical models [14]. The construction of a neural network model requires a large number of samples for training, and although a large number of samples can help improve the accuracy of the model, the actual number of water quality sampling points is affected by weather conditions, cost, and other factors, and fewer sample points can be obtained. However, due to factors such as weather and costs, the number of sampling points obtained is relatively small. For instance, if the sample point is located in the pond or lake area, under bad weather conditions, the repeated measurement of the sampling point data completely depends on manual work; multiple repeated measurements also mean that the collected samples need to be processed many times at the same time, so the number of actual water quality sampling points is generally small. In contrast, support vector machines and random forest algorithms are more suitable for analyzing the little number of samples. For example, Park et al. used both artificial neural networks (ANN) and support vector machines (SVM) to predict Chl-a concentrations in Juam and Yeongsan reservoirs, and the result was that the support vector machine model had higher prediction accuracy compared to ANN [15]. By constructing a random forest model, Fang Xinrui et al. estimated the suspended sediment concentration in the river section from Yichang to Chenglingji downstream of the Chinese Three Gorges Dam for each month before and after the dam construction [16]. Overall, the empirical method is relatively simple, has greater potential for development when combined with artificial intelligence techniques, and is one of the common methods currently used for remote sensing retrieval of TSM concentrations.

The study area used in this study is the Nansi Lake region in North China, which is located in the northern part of the Huaihe River Basin, at the junction of the provinces of Jiangsu, Shandong, Henan, and Anhui, and is a shallow macrophytic lake. As an important water storage lake of the East Route of South-to-North Water Transfer Project, it has the functions of flood control, drainage, breeding, tourism, and so forth [17]. The quality of the water in the area has been gradually declining and neglected by the development of local industry and urbanization for some time, so it is meaningful to detect the TSM concentration in the water in the area. Meanwhile, this study uses Hyperion hyperspectral satellite imagery along with field-measured data to conduct remote sensing retrieval studies of TSM concentrations in Nansi Lake. Models such as linear regression, BP neural network, KNN, AdaBoost, RF, GA_RF, and decision tree are constructed using empirical methods. By comparing the retrieval accuracies of different models, the optimal model is used to retrieve the spatial distribution of TSM concentrations.

2. Data and Methods

2.1. Study Area. Nansi Lake is located in the southwest of Shandong Province and belongs to Weishan County, Jining City, Shandong Province, China, with geographical coordinates of 116°34′E–117°21′E, 34°27′N–35°20′N (Figure 1), and belongs to the temperate monsoon climate. It includes Weishan Lake, Zhaoyang Lake, Dushan Lake, and Nanyang Lake, with a total area of 1266 km^2 and an average depth of 1.5 m. It is the largest and best-preserved inland large freshwater shallow macrophytic lake in the area north of the Yellow River. It is an important water source and water storage lake of the East Route of South-to-North Water Transfer Project. Therefore, it is of great significance to use remote sensing to monitor its water quality.

2.2. Remote Sensing Data and Measured Sample Data. This research uses Hyperion hyperspectral remote sensing data. Hyperion is a hyperspectral sensor carried by an Earth

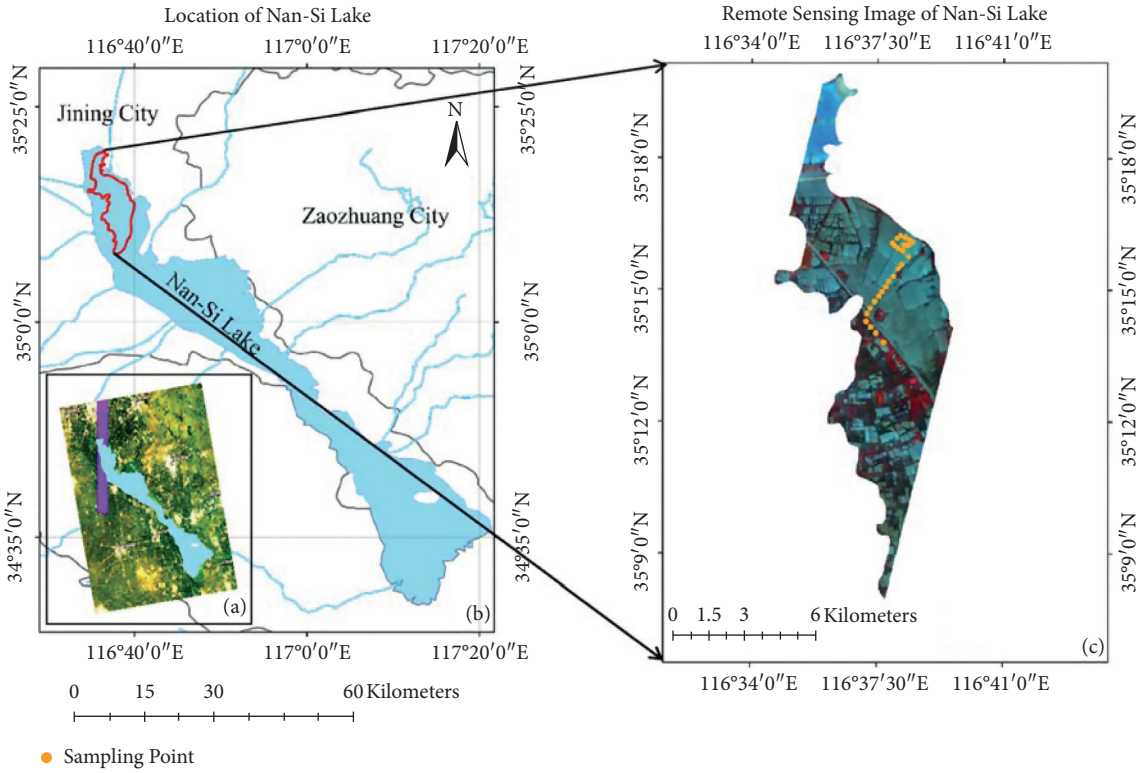

Figure 1: Location and sampling point distribution of study area. (a) Satellite strip map, showing the location of the satellite image within the study area. (b) The location of the study area, as shown by the red mark, which is a typical shallow-grass lake area. (c) Sampling point distribution map, which shows the spatial distribution of sampling points in the study area.

observation satellite named Earth Observing-1 (EO-1). In the range of 355–2577 nm, the Hyperion data provide 242 bands: 35 visible, 35 near-infrared, and 172 short-wave infrared bands. It has a spectral resolution of 10 nm and a spatial resolution of 30 m. A total of 26 sample points were measured in this study, and the data were collected on July 31, 2010, with a clear and cloudless sky, no wind, and a calm water surface. The highest value of TSM concentration at the sample sites was 54.8 mg/L, the lowest value was 4 mg/L, and the mean value was 30.31 mg/L. The image of July 26, 2010, was selected to extract the spectral data of the sampling sites. The imaging date of this hyperspectral image is most suitable with the measured data, and the distribution of sampling points is shown in Figure 1.

3. Methodology

3.1. Backpropagation Algorithm (BP). The backpropagation (BP) neural network algorithm has the characteristics of self-adaptability, strong learning ability, and fault tolerance. BP neural networks combined with remote sensing technology can effectively make a nonlinear prediction of water quality parameters [18].

The BP neural network algorithm is a multilayer feed-forward network trained according to the backpropagation algorithm, which is divided into input layers, hidden layers, and output layers. The core idea is the gradient descent method, which modifies the weight values and threshold values along the negative gradient direction of the function to minimize the difference between the actual output of the network and desired output [19]. The topology of the network structure is shown in Figure 2. The network has two processes: forward propagation of operating signal and backpropagation of error signal. If the difference between the output result and the expected result exceeds the requirement, the error signal will be propagated from the output layer to the input layer, and the weight values and threshold values of each layer of the network will be adjusted to meet the error limitation.

3.2. K-Nearest Neighbors (KNN) Algorithm. The K-nearest neighbor (KNN) algorithm, a theoretically mature method, is widely used in remote sensing quantitative inversion. The sample to be tested can be estimated with its nearest K-neighboring samples, and the attributes of the sample can be obtained by assigning the average of the attributes of the K-neighboring samples to the sample to be tested [20]. The process is described as follows: first, calculate the distance between the sample to be tested and all other samples and then find the K-nearest neighbors of the sample to be tested and finally average the attributes of the selected K-neighbors to obtain the attributes of the sample to be tested [21]. The advantage of using K sample points when estimating pixel-level TSM concentrations using this method is that it reduces random variation due to noise, internal variation, and misalignment of sample point coordinates.

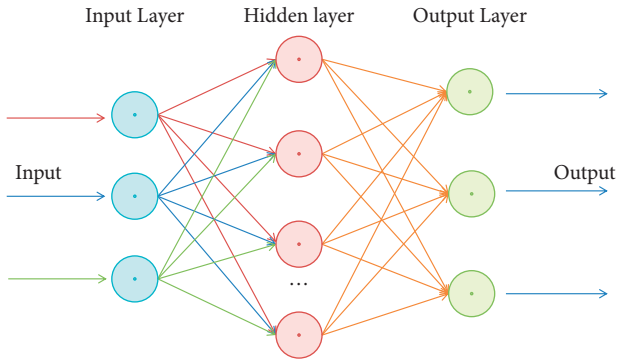

Figure 2: BP neural network structure diagram.

3.3. Decision Tree Regression (Decision Tree). The decision tree regression uses classification and regression trees with the Gini coefficient for constructing a tree. Using the training dataset to construct a decision tree of a certain scale level, the validation dataset is used to prune the constructed tree and select the best subtree that best meets the actual requirements at the cost of losing some information in the pruning process. In this study, the measure of uncertainty of the data at the time of execution of the algorithm is the Gini index, and also the Gini index is used to determine the optimal classification of the variables.

3.4. AdaBoost Algorithm (AdaBoost). The AdaBoost algorithm is a widely used iterative algorithm, as well as a boosting algorithm with adaptive capabilities. The basic idea of this method is as follows: first, strengthen the training of weak classifiers, and then integrate trained weak classifiers to obtain the final classifier with enough strength level. After each iteration, the weights of the samples are adjusted, and the samples with larger fitting errors will increase the corresponding weight values. The weak learner obtains a sequence of functions on the predicted values by iterative operations, and each prediction function is assigned a weight. The function with better prediction results has a larger corresponding weight, and, after several iterations, the final strong learner is obtained by weighting the weak learner function [22]. The main idea is to integrate multiple weak learners to get the output of strong learners to make accurate predictions.

3.5. Random Forest (RF) Algorithm. Random forest was proposed by Leo Breiman in 2001 as a machine learning algorithm formed by combining the ideas of boost aggregation integration and feature random selection [23]. The basic idea of the algorithm is described as follows: the classifier consists of some classification regression trees, and the final result of the classification is decided by voting. Assuming that there are T variables when each tree is split into nodes to generate a decision tree, t ($t < T$) variables are randomly selected as candidates for node splitting, with the same number of variables at each node, and the optimal branches are selected according to the branching optimality criterion. Each tree is recursive from top to bottom until the splitting condition is met. The predicted value of the dependent variable is obtained by averaging the predictions of all trees [24]. The key to the algorithm is to determine the number of variables and the number of decision trees. The algorithm does not overfit as the number of trees increases, has good generalization performance, is more robust, and is suitable for dealing with high-dimensional, nonlinear complex problems [25, 26].

3.6. Optimisation and Selection of Parameters for Genetic Algorithms (GA_RF). The core idea of genetic algorithm comes from the evolution of biology. It is a global probabilistic random search algorithm. It has been widely used in the field of artificial intelligence and has a lot of experience for reference in solving optimization problems. Therefore, the genetic algorithm is adopted in this study to optimize the parameter selection process of the random forest algorithm, and the optimal combination of parameters is found with less manual intervention.

In this study, the optimized algorithm is defined as GA_RF algorithm. The basic idea of the algorithm is to simulate the process of biological genetic evolution. First, the population is initialized, and each chromosome represents a solution. The fitness function measures the quality of the solution and determines the "parents" of the next generation. Then, the next-generation population is generated through crossover and mutation [27, 28]. In the process of genetic algorithm optimization, one chromosome represents a different combination of two parameters of random forest algorithm. After that, the effect of different combinations was evaluated by the fitness function, and the combination method with the best accuracy was finally obtained.

Zhou et al. [29] took the number of features and the number of decision trees as the variables to be optimized in the genetic algorithm and finally evolved into a better parameter combination. This method can ensure computational efficiency and improve the effectiveness of random forest classification. Therefore, this study combines genetic algorithm with random forest algorithm, and the two parameters of optimal design are the number of features and the number of decision trees. In this study, the specific algorithm parameter settings are shown in Table 1.

As shown in Table 1, a certain amount of the initial population is generated within the value range of m and k. After that, the fitness of the initial population is calculated. If the termination condition is not met, a new "population" is generated through selection, crossover, and variogram until the set final condition is reached. In other words, when the "population" evolves to the ideal effect, the optimal solution of the combination of two parameters of the random forest can be output. The advantage of using genetic algorithm to optimize stochastic forest model is that it can reduce manual intervention, avoid local optimal solutions, and can find the global optimal solution in complex space.

4. Experimental Process and Preprocessing

4.1. Experimental Process. Firstly, the Hyperion remote sensing image is preprocessed, and then the spectrum data are extracted from the corresponding sample points, the

Table 1: GA_RF algorithm parameter configuration.

Attribute set	A detailed description
The encoding type	Real number coding
Genetic algorithm chromosome length value	2
Characteristics of the number: m	$1 < m < 32$
Number of decision trees: k	$1 < k < 1000$
Genetic algorithm population size	10
Genetic algorithm mutation rate	0.05
Number of iterations of genetic algorithm evolution	500

spectral characteristics and sensitive bands of the TSM concentration are analyzed, and the high correlation feature variables are screened. On this basis, different methods are used to establish the retrieval model for accuracy evaluation, and, finally, the spatial distribution of the retrieved TSM concentration in the study area was analyzed, and the process of the experiment is shown in Figure 3.

4.2. Data Preprocessing

4.2.1. Removing the Uncalibrated and Vapor-Affected Bands. Among the 242 bands of Hyperion data, 8~57 and 77~224 bands are radiometric calibration bands; the rest of the bands are not calibrated and are set to 0 value. 77~78 bands overlap with 56~57 bands, and the latter is retained due to the large noise, while 121~126, 167~180, and 222~224 bands are influenced by water vapor, which also needs to be removed [30], and the remaining bands are recombined.

4.2.2. Radiation Calibration and Atmospheric Correction. The radiometric correction is to convert the pixel value of the image into an absolute radiometric value (formula (1)). Atmospheric correction is to get the true reflectance of ground objects by eliminating the influence of atmosphere, light, and other environmental factors. The above processes are done in the module corresponding to ENVI5.3.

$$L(k) = C_0(k) + C_1(k) * \mathrm{DN}(k), \tag{1}$$

where k represents the band number and C_0 and C_1 are correction coefficients (offset and gain coefficients, respectively).

4.2.3. Water Information Extraction. Before analyzing the concentration of TSM in the selected waters, the nonwater body information should be removed to focus on the water body information in the study area. The indices popularly used to extract water body information are NDWI [31] (equation (2)) and MNDWI [32] (equation (3)) proposed by Hanqiu Xu using the short-wave infrared band instead of the near-infrared band in NDWI.

$$\mathrm{NDWI} = \frac{G - \mathrm{NIR}}{G + \mathrm{NIR}}, \tag{2}$$

$$\mathrm{MNDWI} = \frac{G - \mathrm{SWIR}}{G + \mathrm{SWIR}}. \tag{3}$$

In the above two equations, G, NIR, and SWIR are green band, near-infrared band, and the short-wave infrared band corresponding to bands 21, 50, and 146 in Hyperion data. The watershed information was extracted and merged using NDWI and MNDWI, respectively. Since the watershed of the study area is a shallow macrophytic lake, the waters are not flat, and the purpose of smoothing its interior is to remove patches.

4.3. TSM Spectral Characteristics and Sensitive Band Analysis.

The atmospherically corrected satellite data is correlated with the components and their contents in the water column [33]. In this study, the measured TSM values are correlated with the spectral reflectance of the corresponding points. The spectral curves of the sampling sites are shown in Figure 4(a). The spectral profile in the visible band is similar to that of vegetation, mainly because the study area is a shallow macrophytic lake with many algae. The reflection peak at 690–780 nm is a typical spectral feature of water bodies containing algae, and the reflection peak at 810–820 nm is formed due to the scattering of light by TSM. As the concentration of TSM increases, the position of the reflection peak shifts towards the long wave direction. The Pearson correlation coefficients for the atmospherically corrected reflectance and TSM concentration at each band are shown in Figure 4(b). They are negatively correlated with 508 nm with a correlation coefficient of −0.41609, positively correlated with 1124 nm with a correlation coefficient of 0.42522, and positively correlated with 844 nm with a correlation coefficient of 0.4211.

In addition, single bands were combined to analyze their correlation with TSM concentrations:

(1) The band difference algorithm: the reflectance data of any two bands are subjected to a different operation ($B1 - B2$). Since the correlation coefficients of the band difference calculation and the differential calculation are consistent, the results of the two algorithms are comparable [34].

(2) Band ratio algorithm: the reflectance of any two bands is calculated by the ratio ($B1/B2$).

(3) Band normalized difference index (NDI) algorithm: the ratio of the difference of any two bands to the sum $(B1 - B2)/(B1 + B2)$.

In this study, combined with the spectral characteristics and correlation analysis of suspended solids concentration, as shown in Table 2, 32 single band and band combination

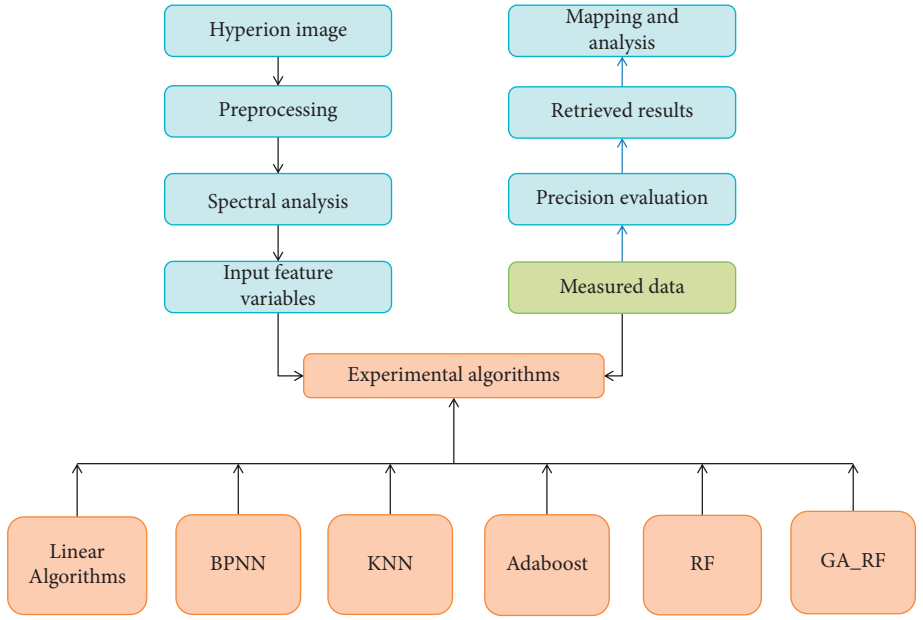

Figure 3: Flowchart for TSM concentration retrieved.

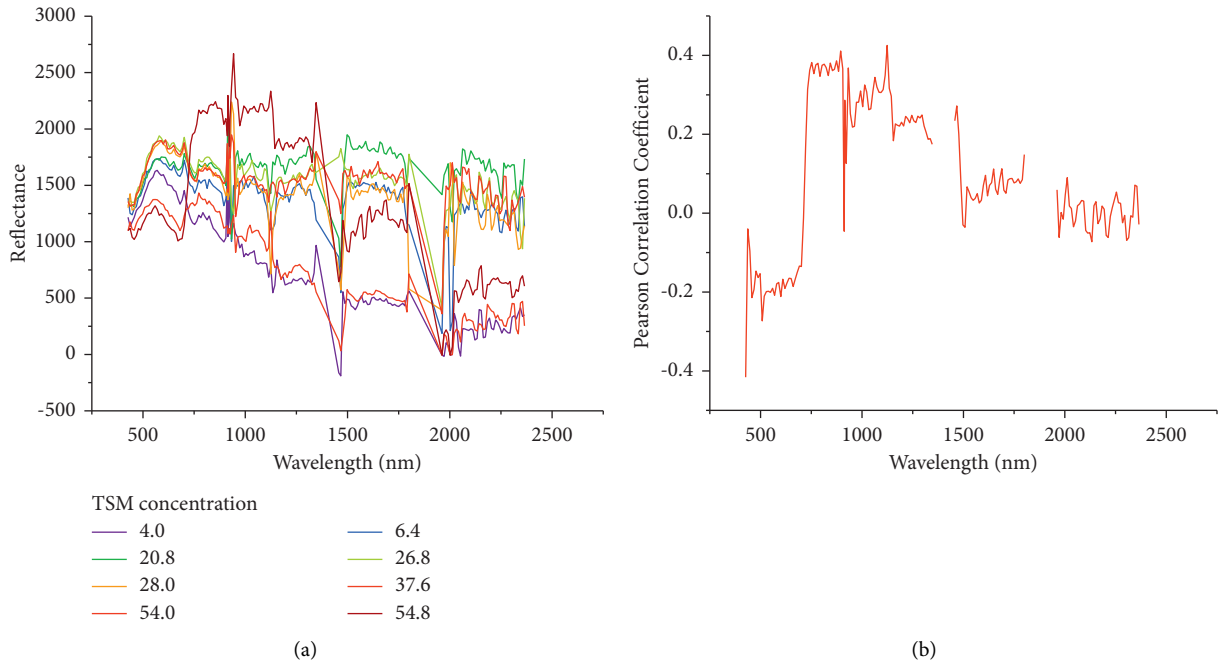

Figure 4: Spectral curves and correlation coefficients of sampling points.

Table 2: Feature vector.

Feature vector	Pearson's correlation coefficient
Band difference algorithm	
(498.04–508.22) nm	0.5527
(2092.84–2133.24) nm	0.5773
(528.57–569.27) nm	**0.6112**
(1648.90–2213.92) nm	**0.6248**

Table 2: Continued.

Feature vector	Pearson's correlation coefficient
Band ratio algorithm	
(1457.23/1991.96) nm	0.5485
(2082.75/2102.94) nm	0.5497
(498.04/508.22) nm	0.5945
(528.57/569.27) nm	0.6076
Band NDI algorithm	
(2082.75–2123.13)/(2082.75 + 2123.13) nm	0.5614
(498.04–508.22)/(498.04 + 508.22) nm	0.5926
(528.57–569.27)/(528.57 + 569.27) nm	0.6074

results are selected as the eigenvectors of the subsequently suspended solids concentration inversion algorithm.

The results show that the highest correlation bands of the difference algorithm are located at 1648.90 nm and 2213.92 nm, while the highest correlation bands of the ratio and NDI algorithms are located at 528.57 nm and 569.27 nm, and the distribution of the highest correlation bands also varies greatly. Combining features of different wave combinations for the model prediction can increase the depth of data extraction.

4.4. Constructing Inverse Models. Based on the latitude and longitude locations of the 26 sampling points, the corresponding values in the reflectance image were extracted and formed into data pairs for inverse modeling and accuracy evaluation of TSM concentration. To reduce the matching error between the sampling point and the image pixel, it is required that the change rate of all pixels in the 3×3 neighborhood centered on the image pixel corresponding to the sampling point is less than 40% [35].

This paper adopts the foldout method [36] for validation: 6 random seeds are set and 6 sets of random numbers are generated for randomly selecting 6 training-testing datasets, where the training dataset contains 20 sample points (77% of the total sample) and the testing dataset contains 6 sample points (23% of the total sample). These six training-testing datasets were used to train and test every one of the models detailed in the following paragraphs. The averages of R^2, RMSE, and ARE from the six training and testing sessions were used as the final model accuracy evaluation metrics for each model.

4.4.1. Linear Models. After the Pearson correlation analysis and significance test of single band divided by combination band, the results are shown in Table 1. Among them, 1648.90 nm–2213.92 nm has the highest correlation. A single-band inversion model and a binary linear model are established. The model and accuracy are shown in Table 3.

The results show that the one regression or multiple regression model performs the worst, since the model only considers the linear relationship between the independent and dependent variables, while the optical properties of Class II waters are complex and the relationship between TSM concentration and the spectrum cannot be fully expressed in linear terms [37].

4.4.2. Machine Learning Models. Machine learning inversion models are constructed with the filtered eigenvector bands as input and the measured total suspended matter concentration as output. The model structure of the neural network is determined by several calculations, and a 3-layer neural network with a network structure of 32-6-1 is used. Due to the small number of samples, the Brute algorithm is chosen for the search algorithm of KNN model. The Gini method is chosen as the feature selection algorithm for decision tree regression models, and the Best method is used to select the feature division points when the sample points are small. The parameters of the AdaBoost algorithm model include base_estimator, n_estimators, learning_rate, and loss. This algorithm is relatively simple, more efficient, and less likely to overfit. The main parameters of the random forest model are the number of decision trees and the number of feature vectors, and the rest of the parameters are set to default values. The specific parameter settings of the above machine learning models are shown in Table 4.

The experimental results are shown in Figure 5. The comparison reveals that the random forest algorithm has the highest accuracy. As a new type of integrated model, it has the characteristics of small training sample requirements and less manual settings of parameters. However, the settings of two parameters of random forest, decision tree, and feature vector have a considerable influence on the results. Therefore, this paper optimizes the selection of its model parameters by introducing genetic algorithm to further improve the accuracy. The main parameters of the genetic algorithm are population size, variation rate, and the maximum number of genetic generations, and the tournament method is introduced to select the next generation. To slow down the convergence rate of the genetic algorithm, the parameters are set as shown in Table 3, taking into account the computing efficiency and effect.

The random forest algorithm's parameters provide a range within which different combinations are made. The experiment of the GA_RF algorithm concludes that when k is certain, the larger m is, the higher the accuracy is; when m is certain, the larger k is, the higher the accuracy is. However, it is not the case that the larger the two parameters are, the higher the accuracy is. After several experiments, the model accuracy regression is higher and stable when $k \in [220, 290]$ and $m \in [18, 26]$ according to the robustness of the results. The final test found that the effect was optimal when $k = 260$ and $m = 22$. The comparison of the effectiveness of the above six machine learning algorithm methods is shown in Figure 6.

Table 3: Linear model and accuracy of suspended solids concentration inversion.

Model type	Formula	Feature bands (nm)	R^2	RMSE (mg·L^{-1})	ARE (%)
Single-band model	$y = 4 * 10^7 * X^4 - 1 * 10^8 * X^3 + 2 * 10^8 * X^2 - 1 * 10^8 * X + 2 * 10^7$	X: 1648.90–2213.92	0.466	246.613	668.671
Binary linear model	$y = -3.89 * X_1^2 - 2.02 * 10^{-5} * X_2^2 + 0.00019 * X_1 + 0.12X_2 + 0.18X_1 * X_2 - 0.0005$	X_1: 1648.9–2203.8301 X_2: 2092.8401–2133.24	0.614	8.679	44.233
	$y = 7.43 * X_1^2 + 0.0002 * X_2^2 + 0.00012 * X_1 + 0.02 * X_2 + 0.18 * X_1 * X_2 - 0.0004$	X_1: 1517.83–2203.83 X_2: 1457.23–1991.96	0.637	8.414	45.274

Table 4: Parameter settings of the machine learning models.

Machine learning model	Parameter settings
BPNN model	Hidden layer neuron function: log S-type function Output function: linear function Training function: momentum BP algorithm with variable learning rate Number of iterations: 1000 Learning rate: 0.003
KNN model	Search algorithm: brute-force search algorithm K: 4 The rest of the parameters: default settings
Decision tree model	Criterion: Gini Splitter: best Max_depth: none Min_samples_split: 2 Min_samples_leaf: none The rest of the parameters: default settings
AdaBoost model	Base_estimator: none N_estimators: 60 Learning_rate: 0.85 Loss: linear The rest of the parameters: default settings
RF model	Number of decision trees (k): 100 Number of features (m): 6 The rest of the parameters: default settings
GA_RF model	Population size: 10 Variation rates: 0.05 Maximum number of genetic generations: 500 Range of the number of decision trees: 5 to 1000; step size: 10 Range of number of features: 1 to 32; step size: 1 The rest of the parameters: default settings

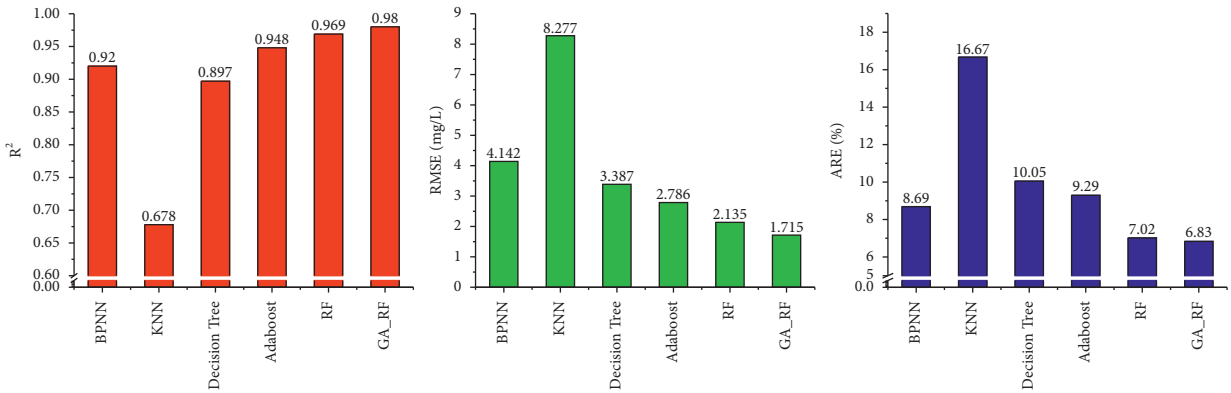

Figure 5: A comparison chart of index data.

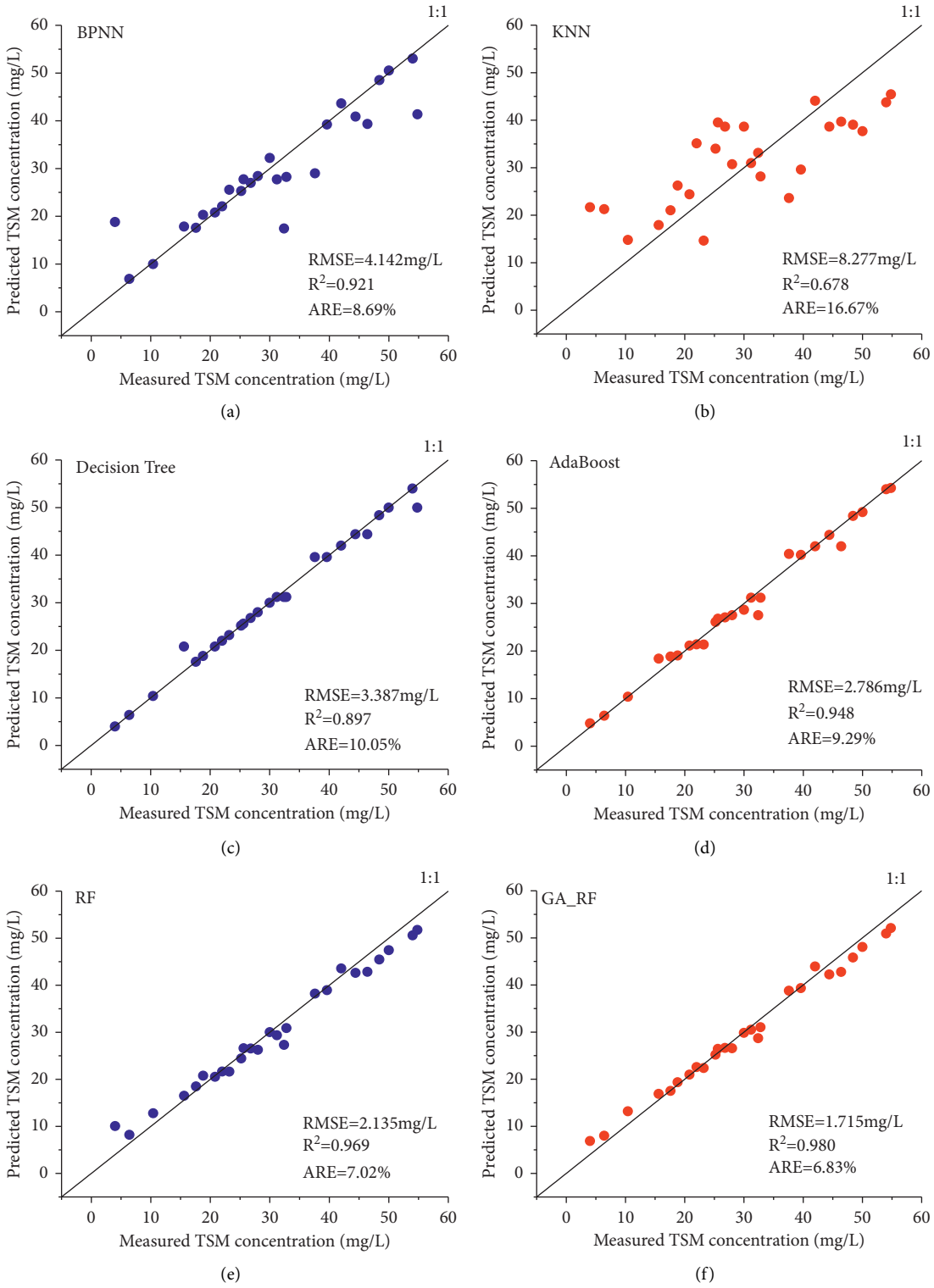

Figure 6: Scatter verification of TSM concentration values estimated by six models and measured values. (a) The BP neural network parameter is a 3-layer neural network. The network structure is 32-6-1, the number of hidden neurons is 6, the number of iterations is 1000, and the learning rate is 0.003. (b) KNN parameters include the search algorithm as brute-force algorithm for violent search, $k = 4$. (c) Decision tree parameters: Gini coefficient is used for feature selection, and the postpruning method is used for the tree correction. (d) AdaBoost parameters: the error calculation function uses the exponential function, and the feature division point method in the weak classifier selects best. (e) RF algorithm parameters: the number of decision trees k is 100 and the number of features m is 6. (f) GA_RF algorithm parameters: the number of decision trees k is 260 and the number of features m is 22.

4.4.3. Spatial Distribution and Analysis of TSM Concentration. According to the above six machine learning models, the TSM concentration was retrieved. The results showed that the concentration of TSM in the study area was higher in the southeast area and lower in the northwest area. However, the spatial distribution of TSM concentration is not uniform. Through analysis, it can be found that the predicted results of different models are not consistent, and the spatial pattern varied greatly between the different results. The predicted values of TSM concentration in KNN and decision tree models are relatively high, while the predicted values of the BP neural network and AdaBoost model are relatively low. The results are shown in Figure 7.

4.4.4. Model Comparison and Evaluation. To verify the prediction accuracy of the models, independent samples were applied as the validation set and several quantitative evaluation indicators were selected to evaluate the accuracy and generalization of the model comprehensively. The model is evaluated using the coefficient of determination (R^2), the root mean square error (RMSE), and the average relative error (ARE) with the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(X_i - Y_i\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{Y}_i\right)^2},$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - Y_i\right)^2}{n}}, \tag{4}$$

$$\text{ARE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{X_i - Y_i}{Y_i} \right| \times 100\%,$$

where Y_i is the measured value; $\overline{Y}_i$ is the mean of the measured value; X_i is the predicted value; and n is the sample size. The larger R^2 is, the smaller the RMSE and ARE indicate higher accuracy and better results. After calculation, the specific operation results of the above model are shown in Table 5.

Through comprehensive comparison and analysis of the above nine models, we found that the accuracy of the linear model is poor, R^2 of the single-band model is less than 0.5, and the accuracy of the binary model has been improved to some extent, but it cannot reach the normal accuracy requirement which is higher than 0.65. The accuracy of the machine learning model has been greatly improved compared with that of the linear model. Among them, R^2 of the BP neural network model is 0.921, and the accuracies of the AdaBoost and RF models are increased by 0.27 and 0.48, respectively, compared with R^2 of the BP neural network. The machine learning model has a better nonlinear fitting and is more suitable for the retrieval of water quality parameters of Class II waters in a complex environment. Among them, the neural network model belongs to the black-box model and has certain restrictions in its use. Although the RF model also belongs to the black-box model, it provides other effective ways to assist the interpretation, making it easier to be explained. In addition, the introduction of two random parameters makes the RF model have better antinoise ability and is less likely to fall into overfitting [16]. The GA_RF results show that R^2 is 0.980, RMSE is 1.715 mg/l, and ARE is 6.83%; and the accuracy of the GA_RF model is further improved based on the RF model.

4.4.5. Further Comparison and Analysis. The random forest algorithm is flexible, robust, simple, and convenient; and it has obvious advantages in parameter optimization, variable ranking, and subsequent variable analysis and interpretation. Using its randomness in selecting samples and independent variables, it can focus on the differences between different samples and independent variables when looking for the relationship between independent variables and dependent variables, so that the regression results can take into account the influence of each sample and independent variables without overtrending to individual samples. The algorithm has a better effect when performing the inversion of water quality parameters with complex change trends such as suspended solids concentration, CDOM.

As shown in Table 6, Silveira Kupssinskü [36] used multiple models to invert the suspended matter concentration using Sentinel-2 imagery and aerial image data obtained by Unmanned Aerial Vehicles (UAVs). The random forest algorithm (RF) eventually yielded the highest accuracy with an R^2 of 0.85603 and an RMSE of 0.00867 m^{-1}. It was more accurate than other machine learning methods. In addition, Wu et al. [38] established an inversion model of CDOM concentration in inland lake waters in China based on Sentinel-3A OLCI data. The results show that the RMSE of the random forest (RF) model is 0.14 m^{-1} and the mean relative error (MRE) is 21%. Compared with the BP neural network model, the RMSE is reduced by 50% and the MRE is reduced by 38%, and the inversion accuracy is significantly improved. In the meantime, Fang et al. [16] constructed different models to inverse the suspended matter concentration based on sediment site monitoring data and MODIS satellite remote sensing reflectance data, and the results showed that the random forest algorithm had higher prediction accuracy and outperformed linear regression, support vector machine, and artificial neural network models.

The above results show that the random forest algorithm has advantages over other machine learning methods in dealing with the inversion of suspended matter concentration in water, which is consistent with the experimental results obtained in this study.

5. Discussion

In this study, the southern part of Nansi Lake in North China was selected as the study area, which is a shallow macrophytic freshwater lake. These lakes are widely distributed, and there are many such lakes with an average depth of fewer than 4 m in the middle and lower reaches of the Yangtze River and Huang-Huai-hai Plain in China, and they are more widely distributed in the world. These lakes are located in inland areas and are important for the local ecological balance, fishery development, and tourism development.

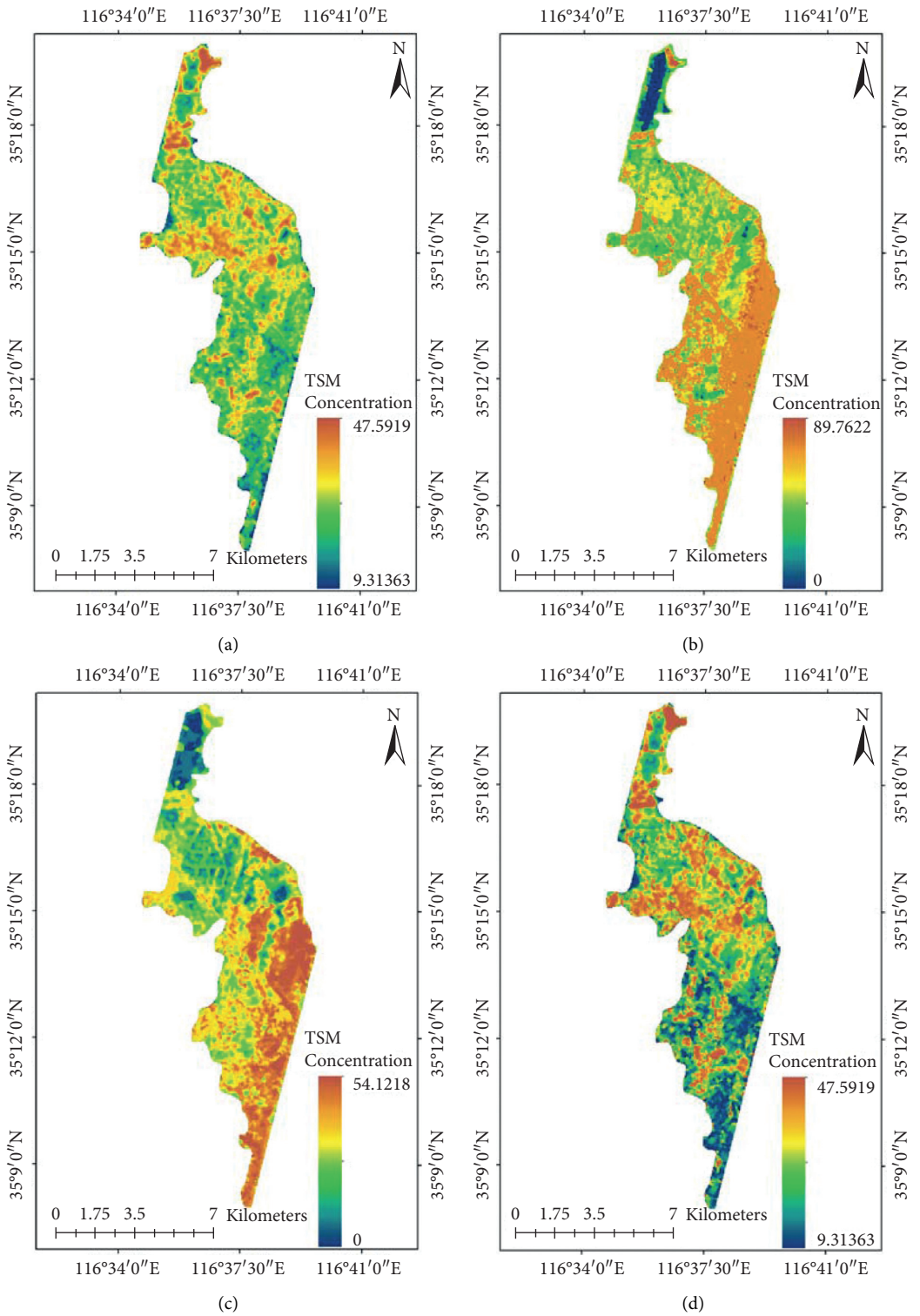

Figure 7: Continued.

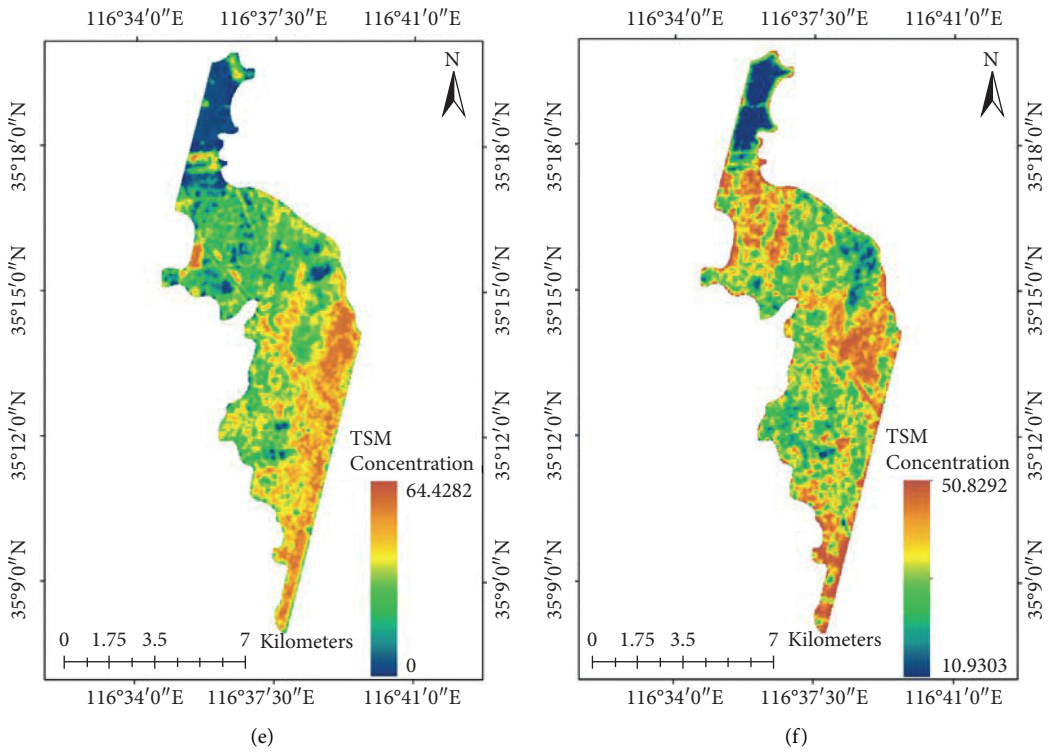

Figure 7: Spatial distribution map of TSM concentration. (a) BPNN. (b) KNN. (c) Decision tree. (d) AdaBoost. (e) RF. (f) GA_RF.

Table 5: Operation results of the models.

Model	R^2	RMSE (mg/L)	ARE (%)
Single-band model	0.466	246.613	668.671
Binary linear model 1	0.614	8.679	44.233
Binary linear model 2	0.637	8.414	45.274
BPNN	0.921	4.142	8.69
KNN	0.678	8.277	16.67
Decision tree	0.897	3.387	10.05
AdaBoost	0.948	2.786	9.29
RF	0.969	2.135	7.02
GA_RF	0.980	1.715	6.83

Table 6: Algorithm test results of other researchers.

	Algorithm name	Evaluation index	
		R^2	MSE
Silveira Kupssinskü et al. [36] conducted experimental comparison results	Linear regression	0.22380	0.04681
	SVR	0.56024	0.02650
	ANN	0.85371	0.00882
	RF	**0.85603**	**0.00867**
	Algorithm name	Evaluation index	
		RMSE (m^{-1})	MRE (%)
Wu et al. [38] conducted experimental comparison results	BPNN	0.24	40
	Single-band model	0.23	36
	RF	**0.14**	**21**

However, the water quality of the shallow macrophytic lake is sensitive, which is easily affected by many factors, such as the surrounding environment, aquatic organisms, and meteorological and geological disasters, as well as aquaculture activities, reclamation, and tourism development. As a result, the inversion of water quality parameters is more

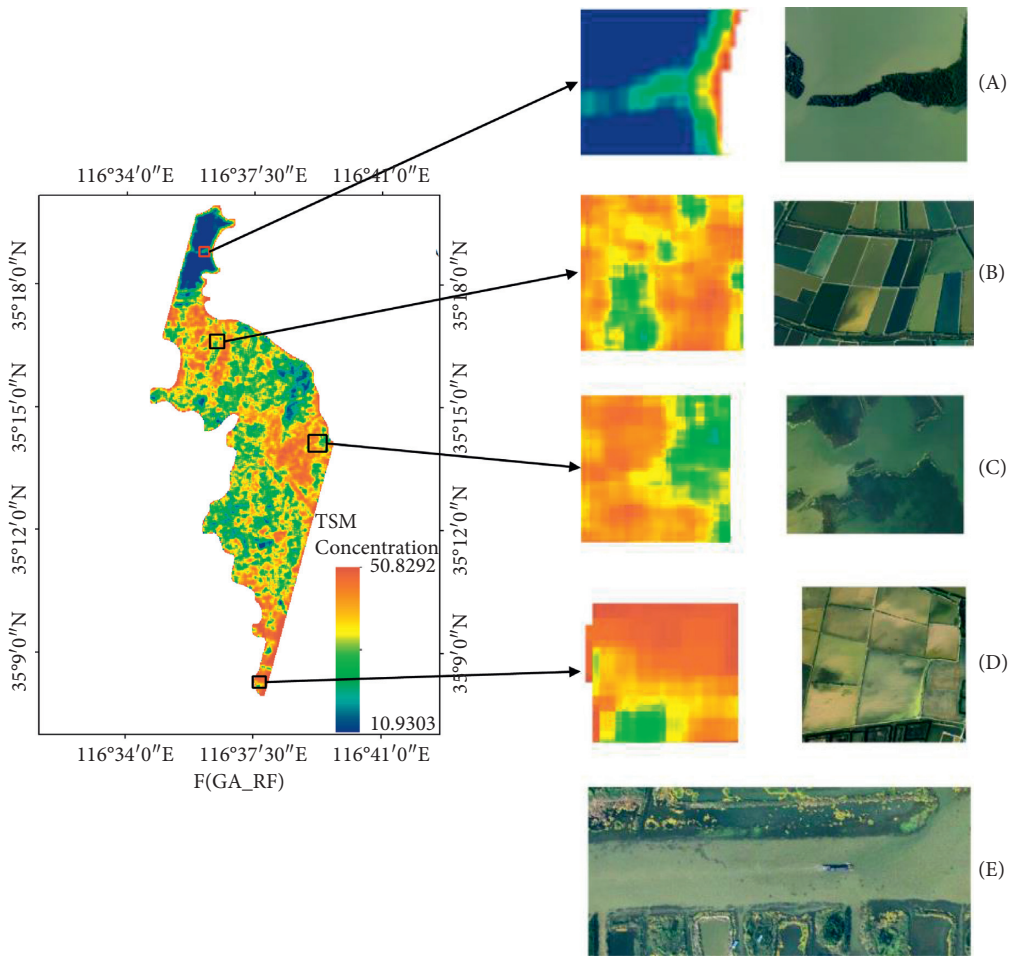

Figure 8: Spatial distribution map of TSM concentration.

complex, and it is difficult to use the conventional inversion model for quantitative monitoring. The ecological problems in the study area selected for this paper are relatively prominent, so this paper proposes using the machine learning method which is more suitable for the shallow macrophytic lake to quantitatively estimate the TSM concentration in the study area and hopes that the conclusions obtained from this study will be of some guidance for the inversion of water quality parameters in other shallow macrophytic lakes in the world.

This study is based on Hyperion hyperspectral data, which have the advantage of high spectral resolution and can better respond to the fine spectral information of the ground material. The spectral analysis shows that the number of bands with high correlation coefficients with the concentration of TSM is small and their correlation coefficients are around 0.4. Therefore, further analysis using band combinations, band difference, ratio, and NDI algorithms yielded higher correlations than single bands, which is consistent with the results of Chinese scholars Xiao et al. [39] and Hou et al. [40]. Some scholars also used the red band to construct the inversion model of TSM concentration [41], and the accuracy was unsatisfactory. Therefore, the univariate factor obtained from spectral analysis and the correlation analysis of the reflectance of each band and the TSM concentration cannot meet the accuracy requirements of the quantitative inversion of TSM concentration in shallow macrophytic lakes. It requires the involvement of multiple types of variables, that is, multiple bands and band combinations jointly for inversion, and this conclusion is consistent with the study by Mao et al. [42]. In this study, 31 feature vectors were selected to participate in the model construction. By comparing and analyzing different algorithms, it can be concluded that the linear model has the worst inversion accuracy, where R^2 of the univariate model is 0.466. The reason for the poor fitting ability of the linear model may be that the linear model can only deal with the linear relationship between the independent and dependent variables, but the TSM concentration parameters are affected by various complex environmental conditions, which often present a nonlinear relationship with high complexity and dimensionality, and even collinearity can occur. This phenomenon is especially true for Class II waters with complex optical properties, where the relationship between water quality parameters and spectra cannot be fully represented using a linear model [34], thus leading to a limited application of linear models. The advantage of machine learning models over linear models is that they can handle the

problem of nonlinear fitting well. These results study that the accuracy of machine learning models such as BPNN, decision tree, AdaBoost, and RF models has been greatly improved compared with that of linear models for the TSM concentration estimation in this study area, with the RF model having the best accuracy and fitting ability. The random forest model is an ensemble model, and it can get higher prediction accuracy than individual learners. Machine learning models such as neural network models and support vector machines have certain shortcomings. Because they are black-box models, it is difficult to understand their internal mechanisms, which restrict their use to some extent. Problems such as overfitting and large calculations will occur at any time. The random forest model is explained by the assistance of each variable on the predictive importance of the model, which some scholars call the gray box model [43]. It has fewer parameters, is simple and easy to use, and has the advantage of being flexible and robust. The random forest model has been widely used because of its high classification accuracy and good robustness, but there is no uniform standard for its parameter setting in academia, and there is no uniform and better method. In this paper, the fit of the model is further improved by introducing a genetic algorithm to further optimize its parameters in the research process. The GA_RF model had the best accuracy with R^2 of 0.980, RMSE of 1.715 mg/L, and ARE of 6.83%. In this study, the spatial distribution of the TSM concentration in the study area predicted by the six machine learning models in Figure 6 is mapped. The linear models are not discussed due to their poor accuracy. The analysis of the mapping results obtained from the six machine learning models shows that there are differences in the spatial distribution of the TSM concentration obtained from different models. The spatial distribution of TSM concentrations predicted by the KNN algorithm in Figure 6(b) shows that the vast majority of the southern and southeastern parts of the study area are high-value areas. The predicted values are high and deviate from the actual situation, failing to effectively predict the TSM concentration in this part of the area. The reason for this may be the small sample size and the failure of KNN to learn its rules effectively. Figure 6(c) shows the inversion results obtained by the decision tree algorithm, which can predict the spatial distribution of suspended matter, but the predicted value is high for the eastern part of the study area. The accuracy of the results obtained from the AdaBoost algorithm shown in Figure 6(d) can meet the requirements. However, there is the problem of poor generalization ability and the overall prediction value for the southern region is relatively low, which is contrary to the real situation. The spatial distribution of TSM concentrations retrieved by the random forest algorithm shown in Figure 6(e) is in better agreement with the actual, but there are also local areas where the predicted values are too high. Overall, the GA_RF algorithm has the highest accuracy and matches the actual situation most closely and can accurately predict the small pollution areas in the region. For example, the northernmost part of the study area is an industrial pollution area, but the local government has prevented the pollution area from spreading through strong and effective environmental protection measures, which makes the pollution area relatively small, and the inversion results can prove that the environmental protection measures implemented by the local government are effective. Therefore, this study concludes that the GA_RF algorithm works best among these algorithms and has an important reference value for water quality parameters' retrieval of shallow macrophytic lakes. It was found by Figure 6 that the TSM concentration showed a trend of high southeast and low northwest, but the spatial distribution of TSM concentration within the region was more fragmented. The lower TSM concentration values in the northern part of the study area may be because the area is the eastern half of Nanyang Lake, located in the scenic area of Taibai Lake, where the water level is deeper and ecological protection is better. The main sources of TSM in the water column include fine-grained sediment carried into the water by surface runoff, phytoplankton, and plant decay residues in the water and resuspension of bottom sediment by wind and wave action [44]. Environmental factors, meteorological factors, and human activities jointly influence the TSM concentration in the study area. The high TSM concentration in the study area is mainly caused by several factors: (1) There are relatively high TSM concentration values in areas such as raised islands in the lake, land, and surrounding shallow water. This situation is shown in Figure 8(a). (2) There are many fish ponds established around the lake in the south and east of the study area, which have shallow depths and high fish densities, and the water bodies suffer from increased disturbances, resulting in increased TSM concentrations, which is consistent with the findings of Meijer et al.'s study [45]. This situation is shown in Figure 8(d). (3) The time of the study was July 31, during which a large number of minnows died in the lake. The flocculent suspended matter formed by the dead residue combined with wind and wave disturbance resulted in an increase in TSM concentration. This situation is shown in Figure 8(c). (4) Frequent human activities, paddling of the lake to make fields (as shown in Figure 8(b)), and the hydrodynamic forces generated by the navigation of many pleasure boats and fishing vessels have caused the resuspension of sediment, as shown in Figure 8(e). (5) The study area is low-lying and many rivers feed into it. When heavy precipitation weather occurs, surrounding rivers carry large amounts of sediment to the confluence, resulting in increased TSM concentrations along the coast. Most of the above factors belong to the characteristics of shallow macrophytic lakes and some other inland lakes, and their water quality parameters are affected by a variety of factors together. The differences in the inversion accuracy of different models are related to not only their algorithms but also the small number of samples, and the accuracy of the atmospheric correction algorithm will have some influence on the prediction accuracy of the models.

There are some weak points in this study. The lake also has different characteristics in different seasons, but only the TSM concentration in a single time phase was inverted in this study. In this study, only spectral factors were analyzed, and the adaptability of the model may be further improved if other influencing factors such as meteorology are combined

for joint retrieval. This work only covers the study area of the Nansi Lake in North China, and further discussion on its applicability to other shallow macrophytic lakes and non-shallow macrophytic lakes is required. Since there are many shallow macrophytic lakes in the world, this study is based on various algorithms for quantitative inversion of TSM concentrations in this type of lake. Comparing the effectiveness of various machine learning methods in analyzing the TSM concentrations in the water environment of these shallow macrophytic lakes, it will provide a more intuitive reference for selecting methods for subsequent research on this issue.

6. Conclusions

Based on the advantage of Hyperion data with the high spectral resolution, the most suitable wavebands and their combinations were selected by spectral analysis for the retrieval of TSM concentration. Using the local area of Nansi Lake as the study area, we quantitatively estimated the TSM concentration using linear models and various machine learning models, verified the accuracy of the inversion data, and mapped the spatial distribution of the TSM concentration. The following conclusions were obtained from the study:

(1) The correlation coefficients of R_{1124} and R_{844} of Hyperion hyperspectral remote sensing data with the concentration of TSM are about 0.42, and the correlation coefficient of $R_{1648.9}$–$R_{2213.9}$ is 0.624. The TSM concentration is influenced by various factors, and there are more variables involved in model estimation. Therefore, TSM concentration inversion cannot use only a certain band but requires the joint participation of multiple variables.

(2) Among the seven types of TSM concentration inversion methods, the linear performed poorly with R^2 of 0.634, RMSE of 8.414 mg/L, and ARE of 45.274%. The accuracy of the machine learning models was greatly improved compared with that of the linear model, in which R^2 of BP neural network, AdaBoost, and RF model were all greater than 0.92 and the RMSE was less than 4.12. The RF model optimized based on genetic algorithm had the highest accuracy with R^2 of 0.98, RMSE of 1.715 mg/L, and ARE of 6.83%. It is concluded that the machine learning models are better than the linear model.

(3) The RF model is simple and robust and has obvious advantages in terms of easy parameter optimization and variable ranking, and the prediction accuracy is significantly better those that of other models, which is more suitable for quantitative remote sensing estimation of TSM in shallow macrophytic lakes.

(4) In the process of actual data processing, the retrieval of TSM concentration in the shallow macrophytic lake is limited by many factors. For example, there are few data sources, low resolution, difficult preprocessing, and few sample points. As a result, the accuracy of the linear model is poor, so the machine learning model is more suitable for remote sensing TSM concentration inversion. For different influencing factors, their characteristics and patterns can be further investigated to quantify their relationship with TSM concentration. The water quality varies greatly in different seasons, and multitemporal sampling point data are required in subsequent studies for better inversion of TSM concentrations. The combination of multiangle analysis of different factors and machine learning models helps to achieve an accurate estimation of water quality parameters by remote sensing methods. Comparing various machine learning methods, it will provide a more intuitive reference for selecting methods for subsequent research on such problems and provide a valuable reference for water quality monitoring and protection of shallow macrophytic lakes.

Data Availability

The data involved in this study can be obtained by contacting the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Xiuyu Liu, Xuehua Li, and Yanyi Li contributed equally to this work. Xiuyu Liu, Xuehua Li, and Yanyi Li carried out the calculation and result analysis and drafted the manuscript, which was revised by all authors. All authors gave their approval of the version submitted for publication.

Acknowledgments

This work was supported by Natural Science Foundation of Shandong Province (no. ZR2019QD010) and Open Fund of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University (no. 19R07).

References

[1] J. Zhao, W. Cao, Z. Xu et al., "Estimation of suspended particulate matter in turbid coastal waters: application to hyperspectral satellite imagery," *Optics Express*, vol. 26, no. 8, pp. 10476–10493, 2018.

[2] Y. Zhang, B. Qin, W. Chen et al., "Experimental study on underwater light intensity and primary productivity caused by variation of total suspended matter," *Advances in Water Science*, vol. 5, pp. 615–620, 2004.

[3] J. Guang, W. Yuchun, J. Huang et al., "Study on seasonal remote sensing estimation model of suspended solids in Taihu lake," *Journal of Lake Sciences*, vol. 3, pp. 241–249, 2007.

[4] D. Pan and R. Ma, "Some key problems in remote sensing of lake water quality," *Journal of Lake Sciences*, vol. 2, pp. 139–144, 2008.

[5] P. Forget, S. Ouillon, F. Lahet, and P. Broche, "Inversion of reflectance spectra of nonchlorophyllous turbid coastal waters," *Remote Sensing of Environment*, vol. 68, no. 3, pp. 264–272, 1999.

[6] J. Chen, T. Cui, Z. Qiu, and C. Lin, "A three-band semi-analytical model for deriving total suspended sediment concentration from HJ-1A/CCD data in turbid coastal waters," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, pp. 1–13, 2014.

[7] S. Buri, Q. Song, and H. Yanling, "Remote sensing retrieval of suspended solids concentration in the Yellow river estuary based on semi analytical method," *Marine Sciences*, vol. 43, no. 12, pp. 17–27, 2019.

[8] V. Klemas, D. Bartlett, W. Philpot, R. Rogers, and L. Reed, "Coastal and estuarine studies with ERTS-1 and skylab," *Remote Sensing of Environment*, vol. 3, no. 3, pp. 153–174, 1974.

[9] F. Gao, Y. Wang, and X. Hu, "Evaluation of the suitability of landsat, MERIS, and MODIS for identifying spatial distribution patterns of total suspended matter from a self-organizing map (SOM) perspective," *Catena*, vol. 172, pp. 699–710, 2018.

[10] X. Hou, L. Feng, H. Duan, X. Chen, D. Sun, and K. Shi, "Fifteen-year monitoring of the turbidity dynamics in large lakes and reservoirs in the middle and lower basin of the Yangtze river, China," *Remote Sensing of Environment*, vol. 190, pp. 107–121, 2017.

[11] S. Lei, J. Xu, Y. Li et al., "An approach for retrieval of horizontal and vertical distribution of total suspended matter concentration from GOCI data over Lake Hongze," *The Science of the Total Environment*, vol. 700, Article ID 134524, 2020.

[12] Z. Zhou, W. Tian, and M. Xin, "Quantitative retrieval of chlorophyll-a concentration by remote sensing in Honghu lake based on landsat8 data," *Journal of Hubei University (Natural Science)*, vol. 39, no. 2, pp. 212–216, 2017.

[13] L. E. Keiner and X.-H. Yan, "A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery," *Remote Sensing of Environment*, vol. 66, no. 2, pp. 153–165, 1998.

[14] J. Chen, W. Quan, T. Cui, and Q. Song, "Estimation of total suspended matter concentration from MODIS data using a neural network model in the China eastern coastal zone," *Estuarine, Coastal and Shelf Science*, vol. 155, pp. 104–113, 2015.

[15] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea," *The Science of the Total Environment*, vol. 502, pp. 31–41, 2015.

[16] X. Fang, Z. Wen, J. Chen et al., "Remote sensing estimation of suspended sediment concentration based on random forest regression model," *National Remote Sensing Bulletin*, vol. 23, no. 4, pp. 756–772, 2019.

[17] C. Yin, Y. Ye, H. Zhao et al., "Analysis of relationship between field hyperspectrum and suspended solid concentration and turbidity of water in Nansi lake," *Water Resources and Power*, vol. 34, no. 1, pp. 40–44, 2016.

[18] H. Lv, X. Li, and K. Cao, "Quantitative retrieval of suspended solid concentration in lake taihu based on BP neural net," *Geomatics and Iformation Science of Wuhan Universty*, vol. 8, pp. 683–735, 2006.

[19] J. Li, J. H. Cheng, J. Y. Shi et al., *Brief Introduction of Back Propagation (BP) Neural Network Algorithm and its Improvement*, Springer, Berlin, Germany, 2012.

[20] Y. Chen, P. Xu, Y. Chu et al., "Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings," *Applied Energy*, vol. 195, pp. 659–670, 2017.

[21] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical Image Analysis*, vol. 35, pp. 475–488, 2017.

[22] H. Reulen and T. Kneib, "Boosting multi-state models," *Lifetime Data Analysis*, vol. 22, no. 2, pp. 241–262, 2016.

[23] L. Breiman, "Using iterated bagging to debias regressions," *Machine Learning*, vol. 45, no. 3, pp. 261–277, 2001.

[24] Y. Qiu, W. Shen, H. Xiao et al., "Water depth inversion based on worldview-2 data and random forest algorithm," *Remote Sensing Ifomation*, vol. 34, no. 2, pp. 75–79, 2019.

[25] Y. Wang, *Optimization and Application of Random Forest Algorithm in Recommender System*, Zhejiang University, Hangzhou, China, 2016.

[26] E. Isaac, K. S. Easwarakumar, and J. Isaac, "Urban landcover classification from multispectral image data using optimized AdaBoosted random forests," *Remote Sensing Letters*, vol. 8, no. 4, pp. 350–359, 2017.

[27] H. Chen, G. Zhao, X. Zhang et al., "Hyperspectral characteristics and content estimation of alkali hydrolyzable nitrogen in fluvo aquic soil based on genetic algorithm and partial least squares," *Chinese Agronomy Bulletin*, vol. 31, no. 2, pp. 209–214, 2015.

[28] A. Yang, J. Ding, Y. Li et al., "Estimation of total phosphorus content in desert soil based on variable selection," *Visible Near Infrared Spectroscopy*, vol. 36, no. 3, pp. 691–696, 2016.

[29] T. Zhou, D. Ming, and R. Zhao, "Parametric optimization of random forest algorithm for land cover classification," *Science Surveying and Mapping*, vol. 42, no. 2, pp. 88–94, 2017.

[30] B. Tan, Z. Li, E. Chen et al., "Preprocessing of EO-1 Hyperion hyperspectral data," *Remote Sensing Information*, vol. 6, pp. 36–41, 2005.

[31] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.

[32] X. U. Hanqiu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 27, no. 12/14, pp. 3025–3033, 2006.

[33] S. Cheng, K. Chang, J. Wang et al., "Discussion on the theoretical model of water color remote sensing," *Journal of Tsinghua University*, vol. 8, pp. 1027–1031, 2002.

[34] F. Yan, S. Wang, Y. Zhou et al., "Study on monitoring water quality of Taihu lake by hyperion spaceborne hyperspectral sensor," *Journal of Infrared and Millimeter Wave*, vol. 6, pp. 460–464, 2006.

[35] E. A. Anas, C. Karem, L. Isabelle et al., "Comparative analysis of four models to estimate chlorophyll-a concentration in case-2 waters using moderate resolution imaging spectroradiometer (MODIS) imagery," *Remote Sensing*, vol. 4, no. 8, pp. 2373–2400, 2012.

[36] L. Silveira Kupssinskü, T. Thomassim Guimarães, E. Menezes de Souza et al., "A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning," *Sensors*, vol. 20, no. 7, 2020.

[37] G. Liu, *Study on the Remote Sensing Estimation Method of Chlorophyll a Concentration Suitable for Different Optical*

Characteristics of Two Types of Water Bodies, Nanjing Normal University, Nanjing, China, 2016.

[38] Z. Wu, J. Li, R. Wang et al., "Remote sensing estimation of colored soluble organic matter (CDOM) concentration in inland lakes based on random forest," *Lake Science*, vol. 30, no. 4, pp. 979–991, 2018.

[39] X. Xiao, J. Xu, D. Zhao et al., "Remote sensing retrieval of suspended solids in typical sections of the middle and lower reaches of Hanjiang river based on hyperspectral data," *Journal of Yangtze River Scientific Research Institute*, vol. 37, no. 11, pp. 141–148, 2020.

[40] X. Hou, F. Lian, H. Duan, X. Chen, D. Sun, and K. Shi, "Fifteen-year monitoring of the turbidity dynamics in large lakes and reservoirs in the middle and lower basin of the Yangtze river, China," *Remote Sensing of Environment*, vol. 190, pp. 107–121, 2017.

[41] K. Shi, Y. Zhang, G. Zhu et al., "Long-term remote monitoring of total suspended matter concentration in lake Taihu using 250 m MODIS-aqua data," *Remote Sensing of Environment*, vol. 164, pp. 43–56, 2015.

[42] Z. Mao, J. Chen, D. Pan, B. Tao, and Q. Zhu, "A regional remote sensing algorithm for total suspended matter in the east China sea," *Remote Sensing of Environment*, vol. 124, pp. 819–831, 2012.

[43] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, 2006.

[44] S. Wang, X. Jiang, W. Wang et al., "Spatiotemporal variation of suspended solids in Lihu lake and its influencing factors," *Environmental Science in China*, vol. 34, no. 6, pp. 1548–1555, 2014.

[45] M. L. Meijer, A. J. P. Raat, and R. W. Doef, "Restoration by biomanipulation of lake bleiswijkse zoom (The Netherlands): first results," *Hydrobiological Bulletin*, vol. 23, no. 1, pp. 49–57, 1989.