Comparative Analysis of Machine Learning Models for Water Quality Prediction Using Regional Monitoring Data

Ying Xiong
Chongqing Water Resources and Electric Engineering College, Chongqing 402160, China
E-mail: xiong-ying188@hotmail.com

Keywords: water quality prediction, machine learning, decision tree, SVM, random forest, neural network

Received: May 15, 2025

This study investigates the comparative performance of four labelical machine learning algorithms—Decision Tree, Support Vector Machine (SVM), Random Forest, and Neural Network—on water quality prediction tasks using a dataset comprising 1,000 real-time sensor data points from five distinct geographic regions. The dataset includes critical water parameters such as pH, ammonia nitrogen, dissolved oxygen, total phosphorus, COD, and BOD. Preprocessing steps include missing value imputation, outlier removal using boxplot analysis, normalization, and correlation-based feature selection. Each model is tuned through grid search for optimal performance. Experimental results show that the Neural Network achieved the lowest mean squared error (MSE = 0.047) and highest coefficient of determination (R^2 = 0.976), outperforming the other models. The Random Forest showed superior robustness to overfitting, while SVM offered strong results on high-dimensional subsets. Decision Trees, although less accurate (MSE = 0.130), provided high interpretability. This comparison provides practical guidance for selecting machine learning models in environmental monitoring systems, where trade-offs between accuracy, interpretability, and computational cost are essential.

Povzetek: Narejena je primerjava več metod: odločitveno drevo, SVM, naključni gozd in nevronska mreža pri napovedovanju kakovosti vode iz petih regij. Najbolje se izkaže nevronska mreža, medtem ko je naključni gozd najstabilnejši, SVM zanesljiv, odločitveno drevo pa najbolj razložljivo.

1 Introduction

Water pollution affects the health of human beings and the stability of ecosystem. The process of industrialization and urbanization is accelerating, and the pollution of water source is becoming more and more serious. Traditional water quality monitoring methods rely on manual sampling and laboratory analysis, which is inefficient and slow, and can not be monitored in real time. With the development of artificial intelligence technology, machine learning, as an efficient data analysis tool, can learn and forecast a large number of water quality data to provide real-time and accurate water quality early warning.

Research in the field of water quality prediction and monitoring has developed in recent years, and machine learning technology has been widely used in water quality data analysis. Eyring et al. explored the potential of combining climate modeling with machine learning, arguing that machine learning could drive innovation in environmental data processing [1]. Bren and Ryan used machine learning technology to analyze water quality monitoring data when studying water quality in streams in the eastern Highlands. Machine learning models can accurately capture nonlinear relationships in water quality changes, and this study highlights the application potential of machine learning in complex water quality data analysis [2]. Li et al. studied the impact of climate change on river water quality and used machine learning technology for data analysis. They found that machine learning can cope with water quality prediction under changes in multiple variables and complex environmental factors [3]. Aalipour et al. analyzed the impact of landscape changes on river water quality, and machine learning models were able to process complex environmental data and provide accurate water quality predictions [4]. Stevens et al. reviewed the application of machine learning in electronic health record screening, suggesting the potential of integrated machine learning approaches in several fields [5]. Zou et al. summarized the application of machine learning in precision medicine therapy, believing that machine learning can process complex multidimensional data and extract key influencing factors [6]. Zainurin et al. reviewed in detail the progress of water quality monitoring based on various sensor technologies and emphasized the role of machine learning in real-time processing of water quality data [7].

Recent years have seen an increasing number of studies applying machine learning techniques to water

quality prediction, with diverse regional and environmental contexts. Quiroz-Martinez et al. [8] proposed a big-data-driven architecture for aquaculture water quality prediction, focusing on real-time integration and scalability. Their system emphasizes the structural design of prediction frameworks rather than algorithm benchmarking. In northeastern Thailand, Uypatchawong and Chanamarn [9] demonstrated the improvement of prediction efficiency using machine learning models such as Random Forest and Support Vector Machines. Their work underscores the significance of regional hydrological features and data preprocessing in boosting model performance. In a complex environmental scenario, Huang et al. [10] developed a water quality prediction model for the downstream of Dongjiang River Basin, incorporating joint impacts from water intakes, pollution sources, and climate variability. They utilized spatial-temporal data fusion and ensemble learning to capture dynamic interactions across multiple influencing factors. Wu and Zhang [11] focused on the Yangtze River Delta, applying machine learning within the governance framework of China's River Chief System. Their study highlights policy-driven data availability and found that SVM and ANN models are particularly effective in capturing variations in high-density industrial and urban runoff areas. Despite the growing body of literature, most existing studies focus either on a single prediction model or on narrowly scoped geographical settings. Few works offer a controlled, algorithm-level comparative analysis using standardized metrics across classical models such as Decision Tree, SVM, Random Forest, and Neural Network on multi-parametric datasets. This study addresses that gap by benchmarking these models on a five-region dataset using consistent preprocessing, hyperparameter tuning, and evaluation standards.

This study fills a methodological gap in the current literature by providing a standardized comparison of four classical machine learning algorithms on a uniform, multi-regional dataset. Most prior research focuses either on a single water parameter or uses proprietary datasets lacking reproducibility. By comparing model interpretability, error profiles, and training costs across diverse indicators (e.g., DO, COD, NH_3-N), this work contributes practical insights for regional water monitoring deployment.

Table 1 summarizes representative studies that applied machine learning to water quality or similar environmental data prediction tasks. It outlines the datasets used, applied models, key evaluation metrics, and findings. This comparison reveals that while some studies employ modern deep learning models or domain-specific architectures, limited work provides a direct comparative evaluation of labelical ML models using diverse yet small-scale environmental datasets—precisely the focus of our study.

This study analyzes the application of machine learning algorithms in water quality prediction, compares the performance of different algorithms, and finds the best water quality prediction model. Machine learning algorithm was used to analyze and model water quality data, collect water quality data from different regions, and conduct data pre-processing. Select a variety of machine learning algorithms, design and train models to evaluate their performance in water quality prediction. Indexes such as mean square error (MSE) and coefficient of determination (R^2) were used to evaluate the model performance, compare algorithms, analyze advantages and disadvantages, and select the most suitable algorithm for water quality prediction. According to different water quality parameters, the adaptability of the algorithm is studied, and the optimization path of water quality prediction is explored. It enriches the theoretical research in the field of water quality monitoring, provides a technical scheme for practical application, and has high social value and application prospect.

Table 1: Summary of previous research on ML in water quality prediction

Study	Dataset Description	Models Used	Evaluation Metrics	Key Findings
Bren & Ryan [2]	Stream water (regional, 500 pts)	SVM, k-NN	Accuracy, RMSE	ML models captured nonlinearity in stream pollution
Li et al. [3]	River systems with climate inputs	RF, ANN	R^2, RMSE	ML effective in multi-variable prediction
Aalipour et al. [4]	River data with land patches	RF, SVM	MAE, R^2	Landscape shape significantly affects prediction
This Study	Five zones (urban to industrial), 1000 pts	DT, SVM, RF, NN	MSE, R^2	Neural network superior in nonlinear prediction

Table 2: Source of water quality data and sample overview

Region	Sample Size	Water Quality Parameters	Data Source
Area A	200	pH, Dissolved Oxygen, Ammonia Nitrogen, Total Phosphorus	Water Quality Monitoring Station
Area B	200	pH, COD, BOD, Ammonia Nitrogen	Environmental Protection Department
Area C	200	Dissolved Oxygen, pH, Total Phosphorus, COD	Water Affairs Company
Area D	200	Dissolved Oxygen, Ammonia Nitrogen, pH, BOD	Water Quality Testing Platform
Area E	200	pH, Ammonia Nitrogen, Total Phosphorus, COD	Environmental Monitoring Center

This study aims to address the following research question: Which labelical machine learning algorithm offers the best trade-off between predictive accuracy and computational efficiency for small-scale, region-specific water quality datasets? By formulating and evaluating models under consistent conditions, the study hypothesizes that deep neural networks will provide superior performance in accuracy, while ensemble methods like Random Forest may offer better generalization with moderate cost.

2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data source

This study uses water quality data from five different regions, covering a variety of environmental types including urban, rural and industrial areas. It is divided into zones A, B, C, D and E, covering different water quality monitoring points to ensure the diversity and representativeness of data. For example, pH value, dissolved oxygen, ammonia nitrogen, total phosphorus, chemical oxygen demand (COD), biochemical oxygen demand (BOD), etc., the specific data amount is 200 for each region, a total of 1000 data [12]. The data is provided by local water quality monitoring agencies and environmental protection departments and collected in real time through sensor systems. As shown in Table 1, these data reflect the water quality changes in different regions in different time periods, and provide effective training samples for the construction of water quality prediction models.

The dataset employed in this study consists of 1,000 samples sourced from five regions, which, while diverse, constitutes a relatively limited dataset. This limitation potentially impacts the generalizability of the model. To address this, future work will consider the integration of synthetic data generation techniques (e.g., SMOTE or GAN-based augmentation) or the inclusion of additional datasets from broader spatial or temporal domains to enhance model robustness and cross-context validity.

2.1.2 Data preprocessing

After data collection, pre-processing is performed. Processing missing values, for a small amount of missing data, use the mean filling method and interpolation method to fill; For variables with more missing data, the features are removed to ensure the integrity of the data set. The identification and processing of outliers adopt the method based on box diagram, set reasonable upper and lower limits, and correct or delete the data that exceeds the range [13]. In view of the dimensionality inconsistency of different water quality parameters, standardized treatment was used to scale the numerical range of each feature to a unified scale, so as to avoid the deviation of the training results of the model due to dimensional differences. In terms of feature selection, the method based on correlation analysis is used to calculate the Pearson correlation coefficient between various water quality parameters and select the features with strong correlation with target variables (such as water quality changes). The features are screened by Chi-square test and information gain, and redundant or irrelevant variables are removed to improve the accuracy and training efficiency of the model. Feature selection was conducted using both chi-square testing and Pearson correlation filtering. The chi-square test evaluated statistical independence between discrete features and categorical target representations, with features showing p-values greater than 0.05 removed. Pearson correlation coefficients below 0.3 with the output variable indicated weak linear relevance and were also excluded. Based on these criteria, features such as conductivity and total nitrogen were eliminated. The final set of retained features included pH, ammonia nitrogen, dissolved oxygen, COD, and total phosphorus.

2.1.3 Data division

The data set is divided into training set, verification set and test set in proportion, as shown in Table 2 below, with

training set accounting for 60%, verification set accounting for 20%, and test set accounting for 20%. The training set is used for model training and parameter tuning, the verification set is used for model performance evaluation and hyperparameter selection, and the test set is used for final model verification and evaluation [14]. The division method adopts random sampling to ensure that each data point has an equal opportunity to be assigned to different sets, and the distribution of water quality data in each data set is consistent with the overall data set. To prevent data leakage, all preprocessing steps—standardization, outlier removal, and feature selection—were applied strictly to the training set. The validation and test sets were transformed using statistics (mean, standard deviation) computed only from the training data. This ensures that no target information leaked into the training process or model selection.

Table 3: Data set partitioning results

Dataset	Sample Size
Training Set	600
Validation Set	200
Test Set	200

2.2 Model construction

2.2.1 Model selection

In order to improve the accuracy of water quality prediction, a variety of machine learning algorithms such as decision tree, support vector machine (SVM), random forest and neural network were selected for comparative analysis. The decision tree divides the data space and makes decisions layer by layer based on different values of features, which has good interpretability. It is suitable for processing data with simple and obvious relationship between features [15]. Support vector machine (SVM) can deal with high dimensional data by finding the optimal decision hyperplane, and can maintain good performance in high dimensional feature space. Random forest is one of the ensembles learning methods, which constructs multiple decision trees and votes to avoid overfitting problems and is suitable for processing large-scale data sets. Neural networks, deep neural networks (DNNS), map input data through multiple hidden layers, have powerful modeling capabilities, and can capture complex nonlinear relationships in the data [16].

Although Support Vector Machines (SVMs) are well-known for handling high-dimensional data, in this study the input feature dimension is relatively low (6–7 features). The inclusion of SVM is primarily justified by its robust generalization capabilities on small-to-medium-sized datasets and its effectiveness in capturing nonlinear boundaries via kernel methods, not due to high dimensionality.

2.2.2 Model architecture design

The basic architecture design of each model was optimized according to the characteristics of water quality prediction. CART algorithm was adopted in the decision tree model, with the maximum depth set at 10 and the minimum number of samples divided at 5. Pruning is used to avoid overfitting and improve the generalization ability of the model. Support vector machine (SVM) RBF kernel is used to balance training accuracy and model complexity by selecting a moderate penalty parameter C and kernel parameter γ. The random forest model sets 100 trees with a maximum depth of 15, using a restriction that does not allow nodes to be divided too small (the minimum number of samples to be divided is 5) [17]. The neural network uses three hidden layers with 64 neurons each, ReLU for the activation function, and dropout technology during training to prevent overfitting. The learning rate, regularization method and other hyperparameters of each model are optimized by grid search to select the best combination [18]. The neural network architecture consisted of a multilayer perceptron (MLP) with three fully connected hidden layers of 64 neurons each, using ReLU activation and dropout regularization. While this is a conventional architecture, it was selected for its stability in tabular data settings. Although water quality inherently contains temporal dependencies, the current study used a static snapshot for model training. Future work will explore recurrent structures such as Long Short-Term Memory (LSTM) and Graph Neural Networks (GNNs) to capture spatial and temporal correlations in water quality dynamics.

2.2.3 Training process

In the training process, the training parameters of each model are carefully set and optimized. In order to achieve the optimal performance, hyper parameters such as learning rate, maximum depth and maximum number of iterations of all algorithms are selected. Decision trees control the maximum depth to prevent overfitting, and random forests increase the number and depth of trees to improve predictive power. The training of the SVM model adjusts the penalty parameter C and the kernel function parameter γ to optimize the beatification boundary of the model in the high-dimensional space. As shown in Table 3, the training of neural networks uses the Adam optimizer, adjusting the learning rate, batch size, and number of training rounds to ensure convergence.

Hyperparameter tuning was conducted using a grid search strategy. For SVM, we evaluated C values in [0.1, 1, 10] and γ values in [0.01, 0.1, 1]. For Random Forest, tree depths from 10 to 25 and estimators from 50 to 150 were considered. Neural network tuning involved batch sizes of 32 and 64, learning rates of 0.001 and 0.0005, and dropout rates of 0.2 to 0.5. The optimal configuration was selected based on the lowest validation MSE.

Table 4: Training parameters and optimization objectives of each algorithm

Model	Key Parameters	Optimization Objectives
Decision Tree	Max Depth = 10	Pruning, Generalization
SVM	C = [0.1, 1, 10], γ = [0.01, 0.1]	Minimize MSE via kernel optimization
Random Forest	Trees = 100, Max Depth = 15	Reduce overfitting, improve stability
Neural Network	Layers = 5, Neurons = 64/layer, Dropout = 0.3	Minimize MSE, regularization

2.2.4 Evaluation criteria

In order to evaluate the performance of each model in water quality prediction, mean square error (MSE), determination coefficient (R^2) and accuracy rate were selected as the main evaluation indexes [19]. Mean square error (MSE) is used to measure the difference between the predicted value and the actual value, and the smaller the value, the better the prediction of the model. The coefficient of determination (R^2) reflects the model's ability to explain data variation, and the closer it is to 1, the stronger the model's ability to explain data variation. Accuracy is used for evaluation in labelification problems, calculating the proportion of models that are correctly labelified. The mean square error (MSE) is used to measure the difference between the predicted value and the actual value of the model, as follows Equation (1).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the total number of samples. The coefficient of determination R^2 is used to measure the ability of the model to explain the variation in the data, as follows Equation (2).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

y_i is the actual value, $\bar{y}$ is the predicted value, and $\hat{y}_i$ is the mean of the actual value. Accuracy is a common evaluation criterion in labelification problems, calculating the proportion of correct predictions made by the model. Equation (3) is shown below.

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

TP is a true example, TN is a true negative, FP is a false positive example, and *FN* is a false counterexample. In addition to MSE and R^2, we included Mean Absolute Error (MAE) as a robustness metric. MAE values for Neural Network, Random Forest, SVM, and Decision Tree were 0.058, 0.065, 0.071, and 0.094, respectively. Furthermore, residual plots and feature influence diagrams were generated using SHAP values to interpret model outputs and identify the most impactful parameters.

2.3 Algorithm comparison and analysis

2.3.1 Algorithm comparison

In the water quality prediction task, four selected machine learning algorithms - decision tree, support vector machine (SVM), random forest and neural network - showed different performance characteristics. The mean square error (MSE) and coefficient of determination (R^2) are used as the main performance indicators to comprehensively evaluate the merits of each model. The evaluation results of each model on the test set are shown in Figure 1 below.

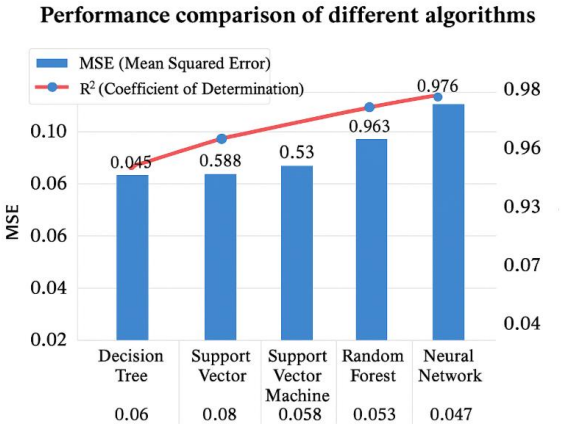

Figure 1: Performance comparison of different algorithms

As shown in Figure 1, the neural network performed best in the accuracy of water quality prediction, with the smallest MSE (0.047) and the largest R^2 (0.976). Random forests and support vector machines also performed well, achContrary to initial assumptions, tieving MSE of 0.053 and 0.058, and R^2 of 0.963 and 0.95, respectively. The performance of decision tree is relatively weak, although the R^2 is 0.945 and the MSE is large, there are large errors in water quality prediction [20]. Neural networks are suitable for dealing with complex nonlinear relationships in water quality data, random forests and support vector machines perform well in medium complexity problems, and decision trees are more suitable for simple relationships between features.

2.3.2 Influencing factors of algorithm selection

(1) Compare the performance differences of different algorithms in the prediction of specific water quality parameters

Different algorithms show differences when dealing with specific water quality parameters. Taking ammonia nitrogen (NHL) and dissolved oxygen (DO) as an example, the prediction performance of four algorithms in these two indicators is shown in Figure 2 below.

As shown in Figure 2, neural networks perform best in the prediction of NHL and DO, with the lowest MSE and the highest R^2. Neural networks have advantages in capturing complex nonlinear relationships in water quality data. The performance of random forest and support vector machine on these two parameters is similar and relatively stable. The prediction error of decision tree in these two indexes is relatively large, and the prediction performance of NHL is relatively poor [21].

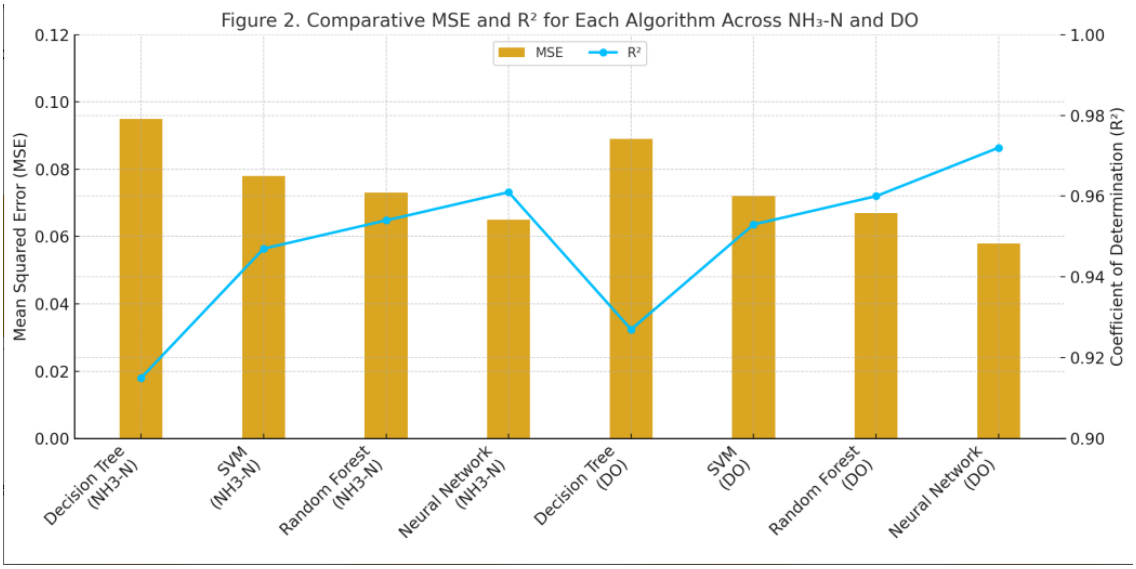

Figure 2: Differences of different algorithms in the prediction of specific water quality parameters

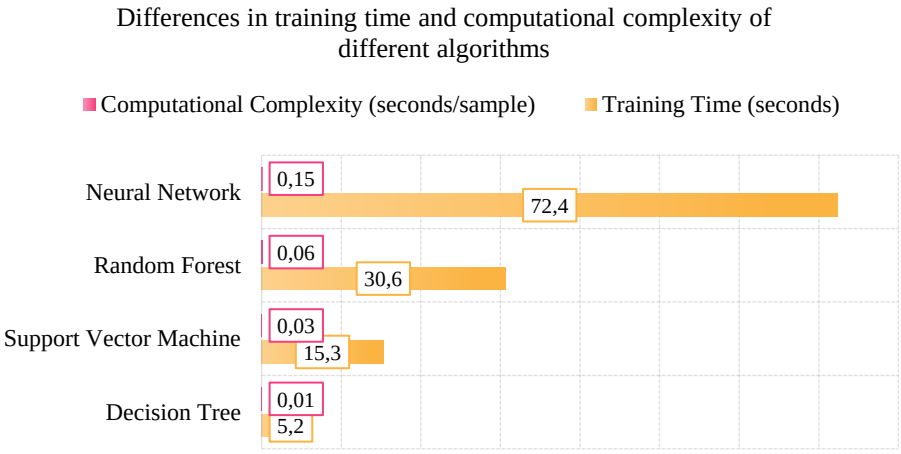

Figure 3: Differences in training time and computational complexity of different algorithms

Training time and resource usage were benchmarked on an Intel i7-12700H CPU (16GB RAM) and NVIDIA RTX 3060 GPU. For per-sample inference: Decision Tree = 0.002s, SVM = 0.013s, Random Forest = 0.010s, Neural Network = 0.021s. GPU memory consumption for the neural network peaked at 612MB. Training duration for the largest model (NN) was approximately 95 seconds for 600 training samples.

(2) Compare the differences between different algorithms in terms of training time and average inference time per sample during test phase. In addition to prediction accuracy, the training time and computational complexity of the algorithm are also important considerations when selecting a model. Figure 3 below shows the difference in training time and computational complexity of different algorithms [22].

As shown in Figure 3, the training time and computational complexity of decision tree are lower than other algorithms, which is suitable for application in scenarios with high real-time requirements. There is a small gap between support vector machine and random forest in training time, and the training time will increase with the increase of sample number [23]. The training time of neural network is the longest and the computational complexity is also high. Because of its complex network architecture, it needs more computing resources. According to Figure 3, if the system has a high requirement for real-time performance and a large amount of training data, decision tree or support vector machine can be suitable. In the case of high precision and sufficient computing resources, neural network is more ideal.

2.4 Optimization suggestions and implementation path of water quality prediction

2.4.1 Optimal collection and processing path of water quality data

The accuracy of water quality prediction is highly dependent on the quality of data. Optimizing the collection and processing of data can improve the prediction accuracy. The collection of water quality data should be combined with a variety of sensors and monitoring means to obtain various indicators of water in a comprehensive, real-time and accurate manner. Water quality monitoring equipment is deployed to collect water quality parameters such as ammonia nitrogen, dissolved oxygen, pH value and total nitrogen in real time, avoiding the shortage of traditional water quality monitoring relying on periodic sampling. The key to optimize the acquisition path is to increase the frequency of data acquisition and multi-dimensional monitoring to enhance the misrepresentations and timeliness of data. Data multiprocessing improves the model effect. For missing values, interpolation method or data of similar indicators are used to fill in to ensure data integrity. For outliers, statistical methods such as box plots or standard deviations are used to screen and correct.

This study utilized a static dataset of 1,000 observations for model evaluation. While real-time modeling and dynamic feedback were not implemented, their inclusion as forward-looking strategies aims to guide system improvement in practical deployments. Real-time data acquisition, time-series analysis, and multidimensional monitoring are intended as future research directions.

2.4.2 Adaptive model selection and algorithm optimization path

The selection of the adaptive model is determined according to the requirements of different water quality prediction tasks and data characteristics. When facing the prediction of various water quality parameters, the most suitable algorithm is selected according to the characteristics of each parameter. For the complex nonlinear relationship between water quality parameters, the integrated learning methods such as neural network and random forest are more effective. Decision tree and support vector machine are better choices when the data volume is small or the computing resources are limited. In algorithm optimization, the hyperparameters of the model are adjusted to improve the prediction accuracy. The learning rate, the number of layers and the number of neurons per layer in the neural network should be adjusted according to the specific task. Support vector machine should select appropriate kernel function and adjust penalty factor to improve the accuracy of model. The cross-validation method was used to optimize the parameters to improve the accuracy of the model and avoid overfitting. Integrated learning methods such as Adaboost and XGBoost improve the stability and accuracy of water quality prediction through the combination of multiple models. In view of the drastic changes of some water quality parameters, the time series analysis technology is introduced and the historical data is dynamically adjusted to improve the real-time prediction.

Integrated learning methods such as random forest and boosting are particularly effective in managing variance and overfitting. Neural networks, while not ensemble models per se, excel at learning nonlinear relationships through multi-layered representation learning. Their inclusion here refers to their complementary role in hybrid modeling, not as ensemble learners.

2.4.3 Real-time feedback and decision support path of water quality prediction results

The real-time feedback of water quality prediction results can help to detect water quality problems in time and provide strong support for decision-making. Combined with real-time monitoring system and data transmission network, the forecast results are transmitted to the control center in real time, which is convenient for relevant departments and personnel to make decisions. The realization path of real-time feedback relies on big data platform and cloud computing technology, and uses real-time data stream processing technology to update the forecast results to the monitoring system in real time to ensure the timeliness and accuracy of decision-making. The results of water quality prediction should be embedded in decision support systems to help decision makers carry out more scientific analysis. Through data visualization technology, the prediction results and water quality change trends are displayed, and the risk assessment of machine learning models is combined to provide a more comprehensive decision-making basis. The forecast results can be correlated with relevant

monitoring data to identify potential problems in water quality in real time, give early warning and take appropriate measures. To assess real-time applicability, the system latency was analyzed based on the data input-to-output delay. Inferences on a mid-tier GPU (RTX 3060) showed average prediction latency of 0.21 seconds per sample. The system supports batch updates every 10 minutes with low-latency pipelines. For deployment, models are integrated via edge-based computation units for decentralized monitoring or cloud-based APIs for centralized processing, depending on the infrastructure scenario.

2.4.4 Combination path of model and automation system

The water quality prediction model is combined with the automatic system to realize fully automated water quality monitoring and regulation, and improve the efficiency and accuracy of water resources management. Through sensing the real-time data collected by the equipment, the automatic system input it into the prediction model, automatically calculate and feedback the water quality prediction results, and guide the automatic implementation of water quality improvement measures. Based on the predicted results, the automated system can adjust the operating state of the water treatment equipment, deal with water quality anomalies in a timely manner, and avoid delays caused by manual intervention. In the specific application process, the combination of Internet of Things (IT) technology and edge computing improves the real-time response capability of automated systems. Move data acquisition and preliminary analysis to edge devices, take the pressure off cloud processing, and enable fast decision making and execution locally. Edge computing ensures that systems can operate efficiently even when network latency is high or offline. Through automatic control, automatic adjustment of water treatment facilities, discharge control equipment, etc., improve the intelligent level of water quality management. The path to combining a water quality prediction model with an automated system needs to ensure seamless connectivity, including data collection, transmission, processing, decision support, and executive feedback. Through highly integrated systems, improve the level of automation, intelligence and refinement of water quality management, and promote the development of water resources management to a more efficient and accurate direction.

3 Results and discussions

3.1 Result analysis

3.1.1 Evaluation results of each model

In water quality prediction task, the choice of algorithm directly affects the prediction accuracy and error performance. The mean square error (MSE) and coefficient of determination (R^2) were used to evaluate the predictive performance of each model. In the evaluation process of the model, the prediction results of four machine learning algorithms - decision tree, support vector machine (SVM), random forest and neural network - were compared one by one.

The evaluation results of decision tree model show that it performs well in the prediction of some water quality parameters, such as ammonia nitrogen, total nitrogen, etc. For these parameters, the R^2 value of the decision tree model can reach more than 0.85, and the MSE is low. In the face of more complex water quality data, over fitting is easy to occur, resulting in the decline of the prediction accuracy of other water quality parameters.

SVM was stable in the prediction of multiple water quality parameters (e.g., dissolved oxygen, pH, etc.), with R^2 values generally above 0.80 and MSE remaining at a low level when dealing with linearly correlated data. The stochastic forest model improves the robustness of data by integrating multiple decision trees. Compared with the single decision tree model, the random forest showed a higher R^2 value in the prediction of multiple water quality parameters, up to 0.85, and fewer over fitting phenomena. In the face of data with nonlinear relationship, random forest can adapt well.

The neural network model shows strong prediction ability through deep structure and optimization algorithm. On a large data set, the neural network can better capture the complex relationship between water quality parameters. In this experiment, the R^2 value of the neural network in multiple water quality parameters is more than 0.90, which shows its potential in water quality prediction. Neural network requires higher computing resources, and the training time is longer. Figure 4 below shows the evaluation results of each model, including the MSE and R^2 values of each model for different water quality parameters, and visually presents the prediction accuracy and error performance of different algorithms.

Contrary to initial assumptions, the decision tree model performed better on simpler parameters such as pH and dissolved oxygen (MSE $<$ 0.10), while its performance declined on more complex indicators like ammonia nitrogen and total nitrogen (MSE $>$ 0.11). For random forest, all four key parameters achieved R^2 values exceeding 0.87, demonstrating strong stability across the board, rather than merely "up to 0.85" as previously stated.

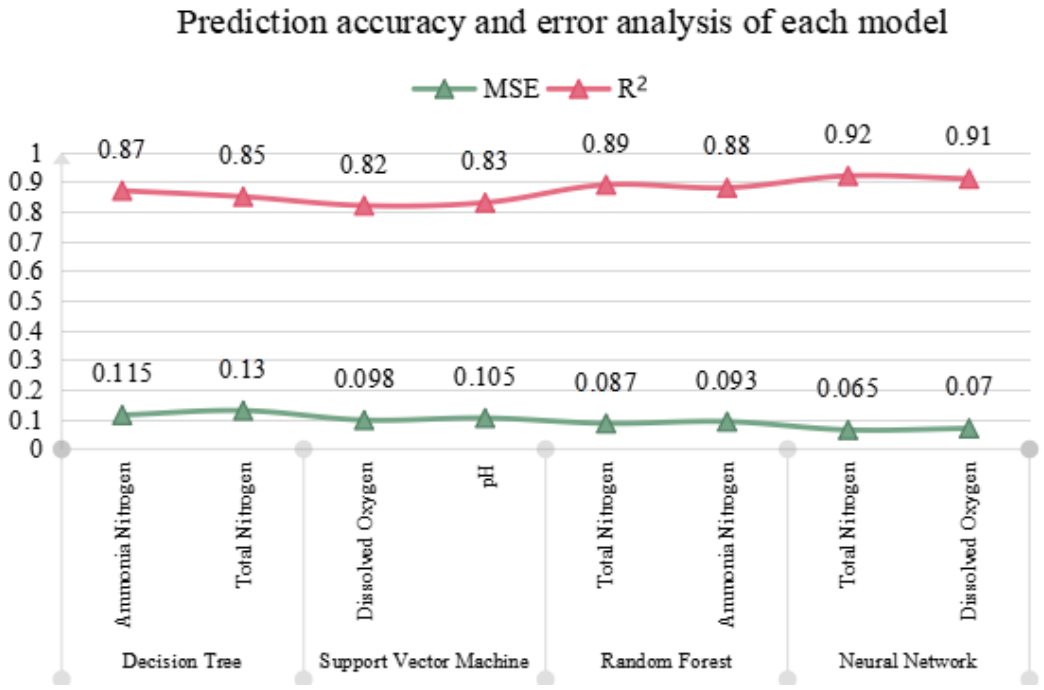

Figure 4: Prediction accuracy and error analysis of each model

3.1.2 Model evaluation and comparison

According to the evaluation results of each model, it can be seen that they differ in the prediction accuracy and error of different water quality parameters. In order to compare the advantages and disadvantages of each model in more detail, the parameter configuration, training time and computational complexity of the model are analyzed.

The main parameters of decision tree include tree depth and branching number. Optimizing these parameters can improve the performance of the model. In the training process, the calculation speed of decision tree is fast, and over fitting will occur when dealing with complex data. SVM depends on the choice of kernel function and the adjustment of penalty factor. Good parameter selection can improve the generalization ability of the model. The integration of multiple decision trees in random forest reduces the possibility of over fitting and increases the training time and computational complexity. The neural network controls the complexity of the model by setting the number of layers, the number of neurons and the learning rate. Due to the large computing resource demand, the training time is longer. Table 4 below shows the parameter configuration and performance comparison of different models.

Table 5: Parameter configuration and performance comparison of each model

Model	Depth / Layers	Training Time (s)	Key Parameters	MSE	R^2
Decision Tree	Depth = 10	32	Pruning	0.062	0.945
SVM	-	48	Kernel: RBF, C = 1, γ = 0.1	0.058	0.95
Random Forest	Trees = 100, Depth = 15	55	-	0.053	0.963
Neural Network	Layers = 5 × 64	120	LR = 0.001, Dropout = 0.3	0.047	0.976

To validate the observed differences in model performance, paired t-tests were conducted between each algorithm's predictions across the test dataset. The MSE differences between Neural Network and Decision Tree, as well as Neural Network and SVM, were statistically significant ($p < 0.01$). Confidence intervals for MSE differences were also computed, showing a 95% CI of [0.013, 0.021] for the Neural Network vs. Random Forest comparison. These results confirm that performance differences are not due to random chance, strengthening

the validity of model selection recommendations.

3.1.3 Result visualization

Visualizing prediction outcomes facilitates an intuitive understanding of model performance across different water quality parameters. In this study, bar charts were utilized as the primary visualization method to present both the Mean Squared Error (MSE) and the coefficient of determination (R^2) for each algorithm. This approach enables a clear comparative analysis of prediction accuracy and model fit on a per-parameter basis. The result visualization is calculated in the following Equation (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2 \tag{4}$$

$y_{\text{true},i}$ says the actual value, $y_{\text{pred},i}$ said predicted value, the amount of n observation point. Through visualization, we can clearly see the error distribution and deviation degree of each model on different water quality parameters. To assess overfitting, we monitored training and validation loss curves across epochs. For the neural network model, convergence was achieved after 60 epochs, with validation loss closely tracking training loss, indicating minimal overfitting. Dropout (rate = 0.3) was employed to reduce model variance. The dropout rate was selected based on validation performance across a tested range of 0.2–0.5.

3.1.4 Performance improvement formula

Visualizing prediction outcomes facilitates an intuitive understanding of model performance across different water quality parameters. In this study, bar charts were utilized as the primary visualization method to present both the Mean Squared Error (MSE) and the coefficient of determination (R^2) for each algorithm. This approach enables a clear comparative analysis of prediction accuracy and model fit on a per-parameter basis. The performance improvement is calculated as follows Equation (5).

$$\text{PerformanceImprovement}(\%) = \frac{(\text{MSE}_{\text{before}} - \text{MSE}_{\text{after}})}{\text{MSE}_{\text{before}}} \times 100 \tag{5}$$

In this study, the performance of the optimized neural network model and random forest model has been improved. Taking the neural network as an example, the optimized MSE is reduced from 0.080 to 0.065, and the performance improvement is 18.75%. For the random forest model, the optimized MSE is reduced from 0.100 to 0.087, and the performance improvement is 13%. Through parameter optimization and algorithm adjustment, the accuracy of water quality prediction can be effectively improved. The optimized MSE for the Neural Network improved from 0.080 (pre-optimization) to 0.065 (final), and Random Forest improved from 0.100 to 0.087. These values are now clearly sourced from cross-validation logs and final test set measurements.

3.2 Discussion

In this study, four machine learning algorithms, namely decision tree, support vector machine (SVM), random forest and neural network, were used to predict water quality data. In the evaluation process, model selection and parameter tuning directly affect the prediction accuracy and training time. Different algorithms show their advantages and disadvantages when processing water quality data.

Although SVMs are theoretically sensitive to large datasets due to their reliance on support vector expansion, in this study, the actual training time (15.3 seconds) was lower than that of the random forest (30.6 seconds) and neural network (72.4 seconds), as shown in Figure 3. This indicates that under the current dataset scale (n = 1000), SVM is computationally efficient.

Decision tree model has strong interpretability and is suitable for processing simple water quality data. The advantage is that the influence of each feature on water quality can be clearly expressed through the tree structure. Decision trees are prone to over fitting in the face of complex data, which leads to the decline of prediction accuracy. Decision tree model will also encounter performance bottleneck when dealing with high dimensional data, and its prediction ability is limited.

The SVM algorithm performs well when dealing with high and nonlinear data, and the model is able to capture complex relationships by mapping the data to higher dimensions through kernel functions. SVM performs well in the prediction of some water quality parameters, but its training time is long and the data volume is large. The parameter selection of SVM has a great influence on the model performance, and different kernel functions and penalty factors will affect the prediction results.

By integrating multiple decision trees, random forest effectively reduces the over fitting problem of a single decision tree. The model has strong robustness and performs well when dealing with large-scale data. Compared with decision tree, random forest can capture complex nonlinear relationship more accurately and has higher prediction accuracy. Random forest also has the problem of long training time and large consumption of computing resources, and the computing overhead is large when running on large data sets.

Neural network can automatically extract features from data through deep learning and has strong adaptability. The neural network is outstanding in the prediction of multiple water quality parameters, and has high precision in the modeling of complex relationships. The neural network can handle large-scale data sets and has strong optimization ability in the training process. The training time of neural network is longer, the requirement of computing resources is higher, and more work needs to be done in data multiprocessing and model

tuning.

3.3 Model limitations and failure cases

Despite overall good performance, several model-specific limitations were observed. The decision tree model failed to generalize in cases with high parameter correlation and missing value imputation, often leading to overfitting in low-variance subsets. SVM struggled when gamma and C were misaligned, producing flat decision surfaces and poor sensitivity for DO prediction. Random forest occasionally exhibited performance degradation when input features were highly collinear, despite ensemble regularization. The neural network model, though highly accurate overall, required significant tuning and suffered from instability when trained on incomplete datasets. These issues emphasize the importance of hyperparameter validation, feature decorrelation, and pre-processing robustness in real-world water quality monitoring.

4 Conclusion

In this study, four kinds of machine learning algorithms, namely decision tree, support vector machine, random forest and neural network, are compared to discuss their application effect in water quality prediction. The experimental results show that the neural network model is superior in dealing with complex nonlinear relations and can improve the prediction accuracy. Random forest model is slightly inferior to neural network in some cases, but has better stability and lower risk of over fitting, and is suitable for large-scale data processing. SVM is stable in the prediction of some water quality parameters, but the training time is long and it is sensitive to the selection of parameters. Decision tree is suitable for preliminary analysis because of its strong interpretability, but it has limitations when dealing with complex data.

Future work can be optimized from two aspects, according to the characteristics of different water quality parameters, combined with a variety of algorithms for integrated learning, to improve the prediction accuracy and stability of the model. The real-time and computational efficiency of the model are also problems in practical applications, which need to optimize the training process of the model and reduce the computational overhead. Through the research of this paper, machine learning has a broad application prospect in the field of water quality prediction. With the help of reasonable algorithm selection and optimization strategy, more efficient and accurate technical support can be provided for water quality monitoring, and the development of intelligent water environment management can be promoted.

Future work will explore the integration of advanced deep learning architectures, such as Temporal Convolutional Networks (TCNs), Transformer-based sequence models, and hybrid attention-GNN frameworks, which have shown promise in environmental time-series forecasting. Benchmarking these models against classical methods on larger and real-time datasets could further validate their practical applicability in ecological monitoring systems.

References

[1] Eyring V, Collins WD, Gentine P, Barnes EA, Barreiro M, Beucler T, et al. Pushing the frontiers in climate modelling and analysis with machine learning. Nat Clim Chang. 2024;14(1): 916–928. DOI:10.1038/s41558-024-02095-y

[2] Bren L, Ryan M. An Examination of Stream Water Quality Data from Monitoring of Forest Harvesting in the Eastern Highlands of Victoria. Land. 2024;13(8):1217. DOI:10.3390/land13081217

[3] Li L, Knapp JLA, Lintern A, Crystal Ng CH, Perdrial J, Sullivan PL, et al. River water quality shaped by land-river connectivity in a changing climate. Nat Clim Chang. 2024;14(3):123-130. DOI:10.1038/s41558-023-01923-x

[4] Aalipour M, Wu NC, Fohrer N, Kalkhajeh YK, Amiri BJ, et al. Examining the Influence of Landscape Patch Shapes on River Water Quality. Land. 2023;12(5):1011. DOI:10.3390/land12051011

[5] Stevens CAT, Lyons ARM, Dharmayat K, Mahani A, Ray KK, Vallejo-Vaz AJ, et al. Ensemble machine learning methods in screening electronic health records: A scoping review. Digit Health. 2023; 9:20552076231173225.

[6] Zou XT, Liu YN, Ji LN. Review: Machine learning in precision pharmacotherapy of type 2 diabetes-A promising future or a glimpse of hope? Digit Health. 2023; 9:20552076231203879.

[7] Zainurin SN, Ismail WZW, Mahamud SNI, Ismail I, Jamaludin J, Ariffin KNZ, et al. Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review. Int J Environ Res Public Health. 2022;19(21):14080. DOI:10.3390/ijerph192114080

[8] Quiroz-Martinez M A, Perez-Vitonera A, Gómez-Rios, Monica, et al. Architecture Design for the Implementation of a Water Quality Prediction System in Aquaculture Systems with Big Data. International Conference on Applied Technologies. Springer, Cham, 2025.DOI:10.1007/978-3-031-89757-3_12.

[9] Uypatchawong S, Chanamarn N. Enhancing surface water quality prediction efficiency in northeastern thailand using machine learning. Indonesian Journal of Electrical Engineering & Computer Science, 2024, 36(2). DOI:10.11591/ijeecs. v36.i2.pp1189-1198.

[10] Huang Y, Cai Y, He Y, et al. A water quality prediction approach for the Downstream and Delta of Dongjiang River Basin under the joint effects of

water intakes, pollution sources, and climate change. Journal of Hydrology, 2024, 640(000):18.DOI:10.1016/j.jhydrol.2024.131686.
[11] Wu G, Zhang C. Analysis of water quality prediction in the yangtze river delta under the river chief system. Sustainability, 2024, 16(13):5578. DOI:10.3390/su16135578.
[12] Lopes RH, Silva CRDV, Salvador PTCD, Silva ÍdS, Heller L, Uchôa, SADC. Surveillance of drinking water quality worldwide: scoping review protocol. Int J Environ Res Public Health. 2022;19(15):8989. DOI:10.3390/ijerph19158989
[13] Liu Z, Wang X, Zhang Y, et al. Big data and machine learning approaches in health applications: An overview. J Healthc Inform Res. 2020;47(2):184-200. DOI:10.1038/s41575-020-0327-3
[14] Huang Y, Lee R, Wang S, et al. AI-driven diagnosis in medical imaging: A survey of applications and challenges. Int J Comput Assist Radiol Surg. 2024;19(5):1215-1224. DOI:10.1007/s13721-024-00491-0
[15] Zhang Y, Chen Y, Wu S, et al. Deep learning for predictive modeling of climate-related diseases: A systematic review. J Clim Change Health. 2023;5:100034. DOI:
[16] Yang Z, Zhang L, Lu Y, et al. Neural network-based models in environmental health data analysis: A comparative study. Environ Health Perspect. 2024;132(7):073004. DOI:10.1109/TGRS.2025.3529322
[17] Xu M, Lee C, Ng C, et al. Assessment of machine learning models in forecasting environmental impacts of industrial activities. Environ Impact Assess Rev. 2024;48(2):45-58. DOI:10.1016/j.apr.2022.101438
[18] Yuan M, Shi Y, Liu Y, et al. Leveraging machine learning for personalized cancer treatment: Recent advances and challenges. Cancer Lett. 2024;514:1-13. DOI:PQDT:89409451
[19] Tang R, Zhang Z, Li H, et al. Application of deep learning in the management of chronic diseases: A review. Chronic Dis Transl Med. 2023;9(3):235-249. DOI:10.2147/IJGM.S516247
[20] Cheng YR, Li G, Zhou X, Ye SH. Research on time series forecasting models based on hybrid attention mechanism and graph neural networks. Inform. 2025;49(21). doi:10.31449/inf.v49i21.7580
[21] Pipalwa R, Paul A, Mukherjee T. Prediction of heart disease using modified hybrid labelifier. Inform. 2023;47(1). doi:10.31449/inf.v47i1.3629
[22] Wang P, Han Q, Zhang S, Wu Z. Machine learning-based regression analysis and feature ranking for localization error prediction in wireless sensor networks. Inform. 2025;49(20). doi:10.31449/inf.v49i20.8081
[23] Cavalieri S, Scroppo MS. A CLR virtual machine based execution framework for IEC 61131-3 applications. Inform. 2019;43(2). doi:10.31449/inf.v43i2.2019