Research Article

Novel PSO Optimized Voting Classifier Approach for Predicting Water Quality

Shweta Agrawal, 1 Sanjiv Kumar Jain, 2 Ajay Khatri, 3 Mohit Agarwal, 4 Anshul Tripathi, 5 and Yu-Chen Hu

1 IAC, SAGE University, Indore, India
2 Medi-Caps University, Indore, India
3 Acropolis Institute of Technology and Research, Indore, India
4 Bennett University, Greater Noida, India
5 UIT RGPV, Computer Science, Bhopal 462 036, India
6 CSIM Providence University, Taichung, Taiwan

Correspondences should be addressed to Shweta Agrawal; shweta_agrawal1883@rediffmail.com and Yu-Chen Hu; ychu@pu.edu.tw

Received 14 June 2022; Revised 29 June 2022; Accepted 1 July 2022; Published 30 July 2022

Academic Editor: Junwei Ma

Copyright © 2022 Shweta Agrawal et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Over the last few years, different contaminants have posed a danger to the quality of the water. Hence modelling and forecasting water quality are very important in the management of water contamination. The paper proposes an ensemble machine learning-based model for assessing water quality. The results of the proposed model are compared with several machine learning models, including k-nearest neighbour, Naïve Bayes, support vector machine, and decision tree. The considered dataset contains seven statistically important parameters: pH, conductivity, dissolved oxygen, Biochemical Oxygen Demand, nitrate, total coliform, and fecal coliform. The water quality index is calculated for assessing water quality. To utilize an ensemble approach, a voting classifier has been designed with hard voting. The highest prediction accuracy of 99.5% of the water quality index is presented by the voting classifier as compared to the prediction accuracy of 99.2%, 90%, 79%, and 99% presented through k-nearest neighbour, Naïve Bayes, support vector machine, and decision tree, respectively. This was further enhanced to 99.74% using particle swarm based optimization.

1. Introduction

Water is life and a vital resource that all living things rely on. It is necessary for all social and economic growth, as well as energy generation, climate change, balancing, and adaptation. It is a huge problem to keep water sources from being contaminated. Humans are the primary cause of increasing levels of many forms of contamination.

All accessible water is not suitable for human consumption, resulting in significant water crises in India. Yet, some low-quality water is available for general use, such as for domestic and industrial reasons. Because of ecosystem degradation, these water sources are being depleted, resulting in widespread contamination of water at deep levels. Water is used for many things, such as drinking water, water supply for homes and businesses, farming, and other human or economic activities. Water quality has a huge influence on both human health and the ecosystem, as drinking dirty water is dangerous when consumed in excess and becomes the reason for diseases such as cholera and diarrhea.

There are a number of interrelated parameters that govern water quality, each of which is influenced by different geographical locations and amounts of water [1]. The majority of research on water quality evaluation applies a range of methodologies, including software analysis that takes into account water flow direction, statistical analysis that makes use of statistics and its tools, and machine learning to identify and assess water quality.
Machine learning-based technologies have been successfully applied in a variety of domains for prediction and identification. Researchers have used machine learning for plant disease detection [2], for pandemic handling [3], cancer diagnostics [4], spondylolysis prediction [5], and stock predictions [6]. There exist many challenges in processing machine learning models [7]. On the other hand, researchers are always looking for new and better ways to improve and optimize existing processes. In the field of machine learning, ensemble learning is an example of a technique that has been proven to produce better results [8]. Using a mechanism such as a majority voting, an ensemble classifier is comprised of a group of individual classifiers combined with a method that aggregates the predictions of the components [9]. According to the research, ensemble classifiers usually outperform classical classifiers in classification tasks. A single base learner algorithm serves as the basis for all members in a homogenous ensemble learning system [10]. However, the organizational structures of the members may differ. The opposite is true of a heterogeneous ensemble, which is comprised of individuals who have a variety of base learners.

To improve classification performance, we offer a homogenous ensemble learning strategy that is both efficient and effective. According to the proposed approach, the dataset is randomly partitioned into smaller groups using a mean-based splitting strategy, and then each division is modelled using the classification and regression tree (CART) algorithm. A comparison with some well-known machine learning methods such as k-nearest neighbour (KNN), Naïve Bayes, support vector machine (SVM), and decision tree (DT) is carried out. Using various machine learning models, several researchers have developed models to simulate and forecast water quality. The feasibility and efficacy of using machine learning applications to forecast drinking water quality have been presented in this research.

Deep learning models to forecast the quality of various water resources such as ponds, rivers, and lakes were presented [11]. The considered quality parameters are level of chlorophyll, turbidity conductance, and level of dissolution. A time-series data-based water quality prediction system for public water utility data is presented in [12]. Different machine learning algorithms like SVM, logistic regression (LR), and deep neural network (DNN) have been used for prediction. A long short term memory (LSTM) based model is presented for water quality prediction.

The deep learning model LSTM has been used by Venkata Vara Prasad et al. [13] to create a water quality prediction system. Seven water quality metrics such as temperature, pH, dissolved oxygen, conductivity, turbidity, CODMn, and NH3–N are selected. A time sequence water condition forecasting system based on the LSTM Neural Network is presented in [14].

Solanki et al. estimated reservoir water quality (WQ). The information was gathered from the Chaskaman reservoir using factors like pH, dissolved oxygen, and turbidity. The performance of the deep learning network ANN in predicting water quality was measured using mean squared error and mean absolute error [15]. Using wavelet neural networks, Xu et al. developed a water condition forecasting model to forecast the water condition of extensive freshwater pearl breeding ponds in Duchang County, Jiangxi Province, China [16].

Haghiab et al. [17] used an artificial neural network (ANN), an association approach of data handling (GMDH), and a support vector machine to estimate the condition of the Tireh River in Iran’s southwest. DO, COD, BOD, EC, pH, temperature, K, Na, and Mg were all evaluated.

Batur and Maktav [18] did another investigation utilizing satellite picture merger and the principal component analysis (PCA) technique to estimate the WQ of Lake Gala (Turkey). Using a decision tree approach, Jaloree et al. [19] sought to estimate the WQ of the Narmada River using five WQ markers. Another study recommended using the deep bidirectional stacked simple recurrent unit (Bi-S-SRU) [20] to develop an accurate WQ forecasting method in smart mariculture.

By combining the ANN and decision tree algorithms, Liao and Sun [21] built a visual to anticipate the WQ of China’s Chao Lake. Yan and Qian [22] presented a least-squares support vector machine-based affinity propagation clustering methodology (AP-LSSVM).

Li et al. [23] used a semantic network and the Markov chain technique to create the latest combination model. Yan et al. [24] proposed using a genetic algorithm (GA) and a particle swarm optimization (PSO) technique to improve the backpropagation (BP) neural network’s ability to estimate the amount of oxygen required in a lake. The prediction findings were found to be more accurate. The authors have mentioned the initialization of position vectors using the weight values of ANN but have not clearly described the upper and lower bound of the values and how it is integrated with the backpropagation weights learning method to update the weights based on PSO or backpropagation. The PSO and genetic algorithm are explained, but not their application for neural network weight optimization. In our proposed work, we have used initial position vector as machine learning parameter values with upper and lower bound as 5 and 20 which can be easily optimized after each training iteration.

Researchers are now focusing on improving the application and authenticity of water condition forecasting/modelling by practicing the latest machinery such as fuzzy logic, stochastic, ANN, and deep learning [25–27].

In forecasting water conditions, Shafi et al. [28] presented four machine learning algorithms: support vector machines, neural networks (NN), deep neural networks, and k-nearest neighbours. To identify the water quality, 25 factors were used as input parameters in single feedforward neural networks [29].

Rankovi et al. [30] adopted the ANN machinery to calculate dissolved oxygen (DO). Gazzaz et al. [31] adopted an ANN model to calculate the water quality index (WQI), and Internet of Things (IoT) technology was used to gather data from water assets. Abyaneh [32] used machine learning techniques such as ANN and regression to forecast chemical oxygen consumption (COD). Sakizadeh [33] estimated the
water quality indicator using ANN with Bayesian regularization. The radial-basis-function (RBF), a kind of ANN machinery, was utilized for water quality forecasting and allocation [33–36].

Mutual information based techniques for input parameter selection are presented and implemented into optimal support vector based regression for seepage driven landslide displacement forecasting [37]. To create a density prediction model and quantify the related predictive uncertainties, a hybrid computational intelligence strategy based on a copula and kernel-based SVM classifier quantile regression was proposed by Junwei Ma et al. [38]. An adapted predictive model is presented for online probabilistic evaluation of voltage security employing decision trees that are updated on a regular basis [39]. The study suggested an ideal combination of approaches for estimating landslide displacement that takes into account the frequency information of deconstructed triggering elements. The observed surface displacements are divided into trend movement and periodic dispersion components during preprocessing using ensemble empirical mode decomposition [40]. Authors proposed a full ensemble empirical mode decomposition for the monitoring displacement of landslides from the three gorges reservoir area and was decomposed into trend and periodic components. The data set of inducing factors was then rebuilt into lower frequency components using the t-test [41].

Chenet al. [42] have calculated WQI using data from rivers in China. Authors have used three ensemble methods, namely, Random Forest, Cascade Random Forest, and Deep Cascade Random Forest. The approach used in our paper differs in the fact that ensemble of different ML models is used which gives high classification accuracy for Indian cities water quality index calculation. The method proposed in [42] uses existing ensemble models, whereas ensemble created by us is entirely new based on KNN, Naive Bayes, SVM, and decision tree. Our paper also introduces a PSO based optimization approach of model parameters to increase the accuracy, which was not present in earlier research.

Partalas et al. [43] have also demonstrated the usage of ensemble machine learning methods for WQI prediction for rivers in Greece. They have used an ensemble of 90 Multilayer Perceptrons and 110 support vector machines. Authors have used different parameter values for both ML methods. Authors have not used any method to fine-tune the parameter values, which is done in our paper using PSO based optimization. Also, the set of ML models used by us are different than those used by authors.

Since the 1990s, ensemble methods have become a prominent learning paradigm. Ensemble learning demonstrated that predictions made by a group of classifiers are often more accurate than predictions made by the best single classifier. Figure 1 shows a simplified depiction. The other is theoretical, in which it has been demonstrated that weak learners can become strong. There is however no much previous research work found on usage of metaheuristics approaches such as PSO for optimization of model parameters. This is the novelty in our proposed approach along with the ensemble of models used in our work, which is not present in existing research work.

1.1. Motivation and Contribution. The motivation for this study regarding water quality predictions is due to the fact that water has a significant impact on both public health and the environment, as drinking unclean water is hazardous and when taken in large quantities, it contributes to disease transmission. The work will be proved helpful to scientists deciding the water quality index using the artificial intelligence with high accuracy. This work can provide them a low cost method to assess the WQI of water, if it is fit for consumption by household or not. Also, the primary goal for this research is to propose and evaluate an alternative technique relying on ensemble machine learning for the precise assessment of water quality. Ensemble learning is the machine learning approach which is proved to yield superior outcomes. Ensemble classifier is typically made up of a set of independent classifiers paired with a strategy that combines the estimates of the elements using a process such like majority voting.

The following are the key contributions of this proposed study:

(1) A novel metaheuristic composite ensemble intelligent algorithm for prediction of water quality index is established.
(2) The work ascertains the development of an efficient voting classifier which is giving high accuracy for water quality detection.
(3) A novel PSO based method is utilised for enhancing the accuracy of voting classifier by optimizing its model parameters. These features are novel and could help researchers to use them for other domains also.

2. Datasets and Parameters

2.1. Dataset Collection. The data for this research was gathered from several historical sites in India. The duration considered for data collection is from 2005 to 2014 and the total collected samples are 1679 from various Indian states. The data is collected from the Indian government website and created by Kaggle.

2.2. Quality Parameters. The collected dataset consists of seven water quality parameters which are pH, conductivity, dissolved oxygen, biochemical oxygen demand, nitrate, total coliform, and fecal coliform.

2.2.1. pH. pH is defined by the negative logarithm of the hydrogen ion concentration. It represents how much acidic or basic property a solution has. The value of pH varies between 0 and 14, where 0 represents the most acidic and 14 represents the most basic.
2.2.2. Conductivity. It is the measurement of water’s capacity to carry electrical current. It depends on the presence of the number of conductive ions in water. These ions are produced by inorganic elements like chlorides, carbonate, and sulphides compounds and salts.

2.2.3. Dissolved Oxygen (DO). This parameter represents the level of oxygen in water resources like rivers, lakes, and streams. It gets reduced due to contamination. Higher values of dissolved oxygen represent better water quality. How much oxygen will dissolve in water is also affected by water temperature.

2.2.4. Biochemical Oxygen Demand (BOD). The quantity of oxygen needed by microbes when decomposing organic materials is known as biochemical oxygen demand. Small levels of oxygen in the form of dissolved oxygen can be found in water. Biochemical oxygen demand is a measurement of the degradation of dissolved oxygen. BOD helps in removing waste organic matter from water.

2.2.5. Nitrate. Naturally, water resources have a moderate value of nitrate. The values beyond permissible limits may be harmful, especially to children. The standard value of nitrate used in public water systems is 10 mg/L.

2.2.6. Total Coliforms. Coliforms are bacteria that are always present in the digestive tracts of animals, including humans, and are found in their wastes. They are also found in plant and soil materials. Total coliform bacteria present in the soil, water affected by surface water, and human or animal waste are all included in total coliforms.

2.2.7. Fecal Coliform. Fecal coliforms are a subset of total coliforms that are thought to be found only in the stomach and feces of warm-blooded animals. Fecal coliforms are regarded as a more accurate indicator of animal or human waste than total coliforms because their sources are more distinct.

2.3. Water Quality Index (WQI) Calculation. Water quality is presented with the help of the water quality index. Normally a WQI is composed of four processes: To begin, the desired parameters for finding the quality of water are identified. Second, the collection of data on identified parameters is done. Thirdly, the weighting factor for each water quality indicator is determined, and fourthly, a final single-value water quality index WQI is produced [39]. WQI has been calculated using the formula shown in equation (1)

\[ WQI = \frac{\sum_{i=1}^{N} q_i \times w_i \sum_{i=1}^{N} w_i }{ } \]  

(1)

N is the considered number of quality parameters of water since there are N parameters involved in the WQI calculations. For each parameter i, equation (2) calculates the quality rating scale qi and equation (3) calculates the unit weight wi for each parameter.

\[ q_i = 100 \times \left( \frac{V_i - V_{ideal}}{S_i - V_{ideal}} \right) \]  

(2)

\[ w_i = \frac{K}{S_i} \]  

(3)

where \( V_i \) is the examined value of parameter i in the tested water samples (some samples are mentioned in Table 1), \( V_{ideal} \) is the ideal value of parameter i in pure water 0 for all parameters except DO = 14.6 mg/l and pH = 7.0, Si is the
recommended standard value of parameter $i$ (as shown in Table 2), and $K$ is a constant. 

Table 2 shows the calculated unit weight.

### 2.4. Influencing Factors

In the water quality index prediction dissolved oxygen, conductivity, fecal coliform, and total coliforms have been considered the most influencing parameters. Higher dissolved oxygen levels indicate higher water quality. Fecal coliforms are believed to be a more accurate predictor of person or animal waste than total coliforms as the variation for the factor, i.e., dissolved oxygen (in mg/l), is from 2.2 to 5.8. The variation of conductivity (in µmhos/cm) is from 39 to 620. Also, the range of deviation for the factor fecal coliform (in MPN/100 ml) lies between 1286 and 11210. The variation range for total coliforms is also very high. All the remaining factors like pH, biochemical oxygen demand, nitrate, and total coliform have a lesser range of variations.

### 2.5. Predicted Content

Water quality index prediction has been done by considering many standard parameters, like pH, conductivity, dissolved oxygen, biochemical oxygen demand, nitrate, total coliform, and fecal coliform. Table 3 shows some calculated water quality indexes.

Table 4 shows the water quality index with the classification of range, where 0–25 states the excellent quality, 26–50 states the good quality, 51–75 states poor quality, 76–100 states very poor quality, and above 100 states not suitable for drinking.

### 3. Methodology

#### 3.1. Algorithms Used

**3.1.1. K-Nearest Neighbour (KNN).** KNN is the initial algorithm in machine learning, and it comes under the scope of supervised learning. It differentiates between the new and presented data and then sorts the data inputs into more suitable grouping in presented data. It suits both classification and regression problems but is customarily used for classification problems. It does not assume substantial data; hence it is nonparametric algorithm.

**3.1.2. Naïve Bayes.** Naïve Bayes is a compilation of classification algorithms based on the Bayes theorem. It is a group of algorithms that all have trivial principles. Naïve Bayes has many types of the algorithm for document classification. Multinomial Naïve Bayes is used and if Boolean is
introduced on this, it turns to Bernoulli Naïve Bayes, and for continuous value predictions, there is Gaussian Naïve Bayes.

3.1.3. Support Vector Machine (SVM). SVM produces momentous accuracy with the minute power of computation. This is the reason every machine learning expert has high trust in it. SVM can be used for both classification and regression problems. For classification, it uses hyperplane to a discrepancy with classes.

3.1.4. Decision Tree (DT). The decision tree is the most famous tool for prediction. It appears like a flowchart where every node is tested to attribute and the outcome is represented by a branch. A leaf node has a class label. The decision tree can deal with the high dimensionality of data and needs very minute computation power. It also deals with both continuous and categorical values.

3.1.5. Logistic Regression (LR). The concept of maximum likelihood estimation underpins logistic regression. The observed data should, in this view, be the most likely. The weighted total of inputs is sent through an activation function that can map values between 0 and 1 in this model. The curve obtained is known as the sigmoid curve or S-curve, and the activation function is known as the sigmoid function.

3.1.6. Random Forest (RF). Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset’s predicted accuracy. Instead of relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. The bigger the number of trees in the forest, the more accurate it is and the problem of overfitting is avoided.

3.1.7. Voting Classifier. A voting classifier is a model which trains other models’ ensembles and delivers a prediction of output based on the highest probability. It intakes the results from various classifiers and then predicts the outcome by more majority. It delivers two types of voting: hard and soft voting. In hard voting, the outcome is based on the majority of voting, while in soft voting, the outcome is based on the average of votes. The reason behind choosing the ensemble learning of people is it has a low error rate and very little overfitting.

To optimize the model, we have used the growing and pruning technique in ensemble learning. Growing involves adding models to the ensemble to improve the accuracy and pruning members from a fully defined ensemble to reduce the model or computational complexity of an ensemble with little or no effect on the performance of an ensemble. Figure 2 shows the model of the voting classifier.

3.1.8. Particle Swarm Optimization. PSO is an optimization process which mimics flock of birds or fishes which optimizes their search for food by learning from each other. Thus a position vector of a population pool having say 100 particles is initialized with random values between an upper and lower bound. These position vectors are updated iteratively based on particle velocity as shown in equation (4)

\[ X_i(t + 1) = X_i(t) + V_i(t + 1). \]  

Here \( X_i \) is the position vector and \( V_i \) is the velocity vector of \( i \)th particle and \( t \) is the iteration number. Similarly the best particle position vector is maintained across all iterations and a global population best position vector is also maintained based on a fitness score. The velocity is updated based on equation (5)

<table>
<thead>
<tr>
<th>Table 3: Calculated water quality index.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total coliform (MPN/100 ml)</td>
</tr>
<tr>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>8391</td>
</tr>
<tr>
<td>5330</td>
</tr>
<tr>
<td>8443</td>
</tr>
<tr>
<td>5500</td>
</tr>
<tr>
<td>4049</td>
</tr>
<tr>
<td>5672</td>
</tr>
<tr>
<td>9423</td>
</tr>
<tr>
<td>4990</td>
</tr>
<tr>
<td>4301</td>
</tr>
<tr>
<td>7817</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: Classification range with WQI.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water quality index</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>0–25</td>
</tr>
<tr>
<td>26–50</td>
</tr>
<tr>
<td>51–75</td>
</tr>
<tr>
<td>76–100</td>
</tr>
<tr>
<td>Above 100</td>
</tr>
</tbody>
</table>
Here \( c_1 \) and \( c_2 \) are constants chosen as 1 and 2 and \( r_1 \) and \( r_2 \) are random values between 0 and 1. Also \( w \) is the inertia factor taken as 0.5; \( p_{\text{best}} \) and \( g_{\text{best}} \) are particles’ best position vectors and global best position vectors depending on fitness score which is the accuracy of the ensemble model.

3.2. Performance Matrices. For the effectiveness of the classifier, there is confusion matrix, which shows True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy: It is the right values predicted by the model on the given input. But accuracy does not provide information about False Positive and False Negative. The accuracy is obtained by True Positive + True Negative divided by True Positive + True Negative + False Positive + False Negative. Equation (6) shows the formula.

\[
\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}
\]

Precision: It is the frequency of a model to predict True Positive. A low precision value leads to a high number of False Positives. It is the True Positives divided by True Positives + False Negatives. Equation (7) shows the formula.

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

Recall: This parameter provides information about the model prediction of False Negatives. It is the True Positives + True Negatives divided by True Positives + False Negatives. Equation (8) shows the formula.

\[
\text{Recall} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN}}
\]

F1-Score: It displays a low number of False Positives and False Negatives when F1-score is high. It is calculated by Precision and Recall: Precision * Recall divided by Precision + Recall multiplied by 2. Equation (9) shows the formula.

\[
\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Support: It is the number of true values that lie in each class of target values.

3.3. Ordering-Based Ensemble Pruning. This work investigates ordering-based ensemble pruning to eliminate redundancy and increase ensemble classification accuracy. It is less computationally costly than search and optimization based ensemble pruning strategies. Ordering-based ensemble pruning sorts all of the classifiers by importance and then prunes the ensemble using some of the top classifiers from the ordered list. The ordering-based ensemble pruning flowchart is presented in Figure 3.

4. Results and Discussion

To reduce the model or computational complexity of an ensemble with little or no effect on the performance of an ensemble we apply the pruning to the ensemble. We removed logistic regression and random forest, from voting classification estimators, and found that there is no difference in the accuracy. So, we removed these two classification models from our ensemble voting model. In our model, we have four base models KNN, Naive Bayes, SVM, and decision tree. We used a voting ensemble with hard voting. When using majority voting with equal weights, the predicted label’s mode is used.

The results of applying various algorithms and techniques like voting, KNN, Naive Bayes, SVM, and
decision tree are determined by many parameters like accuracy, precision, recall, F1-Score, and support. The dataset is divided into 80% for training and 20% for testing samples. Table 5, shows that ensemble learning achieved an accuracy of 99.5%, which is the highest among the accuracy of 99.2%, 90%, 79%, and 99% presented through KNN, Naïve Bayes, SVM, and decision tree, respectively.

Figures 4–8 present the confusion matrix of a decision tree, KNN, NaïveBayes, SVM, and voting classifier, respectively. In the confusion matrix, the deeper the colour of the class, the more accurate the result; similarly, the more
Figure 4: Decision tree based confusion matrix.

Figure 5: KNN based confusion matrix.
Figure 6: Naïve Bayes based confusion matrix.

Figure 7: SVM based confusion matrix.
diagonally oriented is the blocks in the matrix, the more accurate and effective the classifier produces.

4.1. Voting Classifier Optimized Using Particle Swarm Optimization (PSO). The accuracy of the ensemble model with default parameters was 99.49%, with 2 misclassifications but using PSO. The test accuracy is improved to 99.74% with only 1 misclassification.

A novel method was created to find the best set of parameters for the voting classifier, which was an ensemble of SVM, k-NN, DT, and Naïve Bayes. To find the best possible parameters, a metaheuristics based method of PSO was used in which an initial population of particles position vectors was initialized with random values between 5 and 20. This vector length is equal to the number of parameters that were tuned in these ensemble classifiers. PSO fitness function is based on the accuracy of the model, which it tries to maximize. As PSO optimizes, we can get the best set of parameter values giving the best model performance. In PSO, particle position in the next iteration is updated based on particle velocity in the new iteration. The velocity is also updated depending on the particle’s best position vector and the global best position vector of the population pool. The number of particles chosen was 15. The equations (from equation nos. 10–15) for model parameters are shown in Table 6.

Here $x$ is the PSO position vector. If the values go outside the bound they are modified to lower bound if the value is lesser than it and to upper bound, if the value is more than it.

The constant value multiplied or divided to vector values was taken by estimating the original default values and as they change, better accuracy parameter values can be easily obtained. Figure 9 presents the confusion matrix of a PSO optimized voting classifier.

![Confusion Matrix](image-url)

**Figure 8: Voting classifier based confusion matrix.**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Methods</th>
<th>Equations</th>
<th>Equation no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>K-NN</td>
<td>$n_neighbours = x[0]$</td>
<td>(10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$leaf_size = x[1]\times 6$</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$C = x[2]/5.0$</td>
<td>(12)</td>
</tr>
<tr>
<td>2</td>
<td>SVM</td>
<td>$random_state = x[3]$</td>
<td>(13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$min_samples_split = x[4]/20.0$</td>
<td>(14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$min_samples_leaf = x[5]/40.0$</td>
<td>(15)</td>
</tr>
<tr>
<td>3</td>
<td>DT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 6: Equations of different models.**
5. Conclusions

Assessing water quality through machine learning will play a very important role in policymaking and water management. The paper presented an ensemble learning-based machine learning model for predicting water quality. For quality prediction, the water quality index has been calculated by considering many standard parameters. The performance of the ensemble-based voting classifier has been compared with other machine learning models like KNN, Naïve Bayes, SVM, and decision tree. Experimental results reveal that the highest presented accuracy is 99.5% which is through the voting classifier. The voting classifier model further improved using particle swarm optimization and this enhanced the accuracy to 99.74%.

The main contributions in this study are that a novel method is adopted to find the best set of parameters for the voting classifier, which is an ensemble of SVM, k-NN, DT, and Naïve Bayes. The test accuracy is improved to 99.74% using PSO. The limitation of this study is that the proposed algorithms are applied on fixed dataset and not considering the stochastic patterns. In terms of the improvement in this work, which also reflects the future work, it is using the some more latest metaheuristic optimization based ML predictions for different water quality indexes. Also, the present work may be utilized as the reference for further improvements in the water quality predictions utilizing the newest machine learning algorithms.

Data Availability

The data used to support the findings of this study are available from the authors upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


J. Ma, Y. Wang, X. Niu, S. Jiang, and Z. Liu, "A comparative study of mutual information-based input variable selection
strategies for the displacement prediction of seepage-driven landslides using optimized support vector regression,” *Stochastic Environmental Research and Risk Assessment*, pp. 1–21, 2022.


