Environmental Science and Pollution Research (2023) 30:22532–22549
https://doi.org/10.1007/s11356-022-23418-6

RESEARCH ARTICLE

A decision fusion method based on classification models for water quality monitoring

Mohamed Ladjal[1] · Mohamed Bouamar[1] · Youcef Brik[1] · Mohamed Djerioui[1]

Received: 21 November 2021 / Accepted: 29 September 2022 / Published online: 27 October 2022

Abstract

Monitoring of water quality is one of the world's main intentions for countries. Classification techniques based on support vector machines (SVMs) and artificial neural network (ANN) has been widely used in several applications of water research. Water quality assessment with high accuracy and efficiency with innovational approaches permitted us to acquire additional knowledge and information to obtain an intelligent monitoring system. In this paper, we present the use of principal component analysis (PCA) combined with SVM and ANN with decision templates combination data fusion method. PCA was used for features selection from original database. The multi-layer perceptron network (MLP) and the one-against-all strategy for SVM method have been widely used. Decision templates are applied to increase the accuracy of the water quality classification. The specific classification approach was employed to assess the water quality of the Tilesdit dam in Algeria as a study area, defined with a dataset of eight physicochemical parameters collected in the period 2009–2018, such as temperature, pH, electrical conductivity, and turbidity. The selection of the excellent parameters of the used models can be improving the performance of classification process. In order to assess their results, an experiment step using collected dataset corresponding to the accuracy and running time of training and test phases, and robustness to noise, is carried out. Various scenarios are examined in comparative study to obtain the most results of decision step with and without feature selection of the input data. From the results, we found that the integration of SVM and ANN with PCA yields accuracy up than 98%. The combination by decision templates of two classifiers SVM and ANN with PCA yields an accuracy of 99.24% using k-fold cross-validation. The combination data fusion enhanced expressively the results of the proposed monitoring framework that had proven a considerable ability in surface water quality assessment.

Keywords Surface water quality monitoring · Principal component analysis · Feature selection · Support vector machine · Artificial neural network · Decision templates

Introduction

The quality of surface water plays a crucial and strategic role in people's health, sustainable development, and ecological systems (Wang et al. 2013). However, due to its limited availability, freshwater is subject to contamination from various sources, including home and industrial pollutants, agricultural runoff, and other sources (Soltani et al. 2021; Oukil et al. 2021). The deterioration of freshwater quality currently is one of the biggest environmental issues (Dilmi and Ladjal 2021). Surface freshwater resources, such as rivers, lakes, and reservoirs, are also important and require careful treatment because the underground water supply is generally insufficient to meet market demand (Soltani et al. 2020). Because it can be used immediately and does not require costly treatments or, more crucially, pose a danger

Responsible Editor: Xianliang Yi

Highlights

- New intelligent water quality classification is performed using data combination fusion and features selection.
- ANN and SVM methods have been proposed for water quality classification status.
- The final decision is performed using decision templates rule combination based on probabilistic output from both the two classifiers.
- Real database from Tilesdit dam (Algeria) are used for evaluation.
- A superior accuracy of up to 99.24% was obtained by the proposed approach and of all used methods.

✉ Mohamed Ladjal
mohamed.ladjal@univ-msila.dz

Extended author information available on the last page of the article

of disease, groundwater is typically selected as a source of freshwater. Efficient information about the location and quality of surface water and groundwater helps for approving the performances and assessment in diverse scientific fields, such as the surface water survey and management, water resource assessment, and monitoring and environmental pollution (Zhou and Wu 2008). Water quality control and monitoring play an important role in the ecological running management, and it presents a considerable concern for conservation and rational utilization of water supplies in the world (Bouamar and Ladjal 2012). The most generally used criteria for evaluating and monitoring water quality convey ecosystem health, public safety, the level of water pollution, and the quality of drinking water. Water quality parameters are determined by the intended usage. Water that has been treated for potability, industrial/domestic use, or restoration (of an environment/ecosystem, typically for the health of people/aquatic life) is the main subject of water quality study. Thresholds, usually known as guidelines, have been defined on selected parameters in order to measure the quality of water for drinking, irrigation, and other purposes. The quality of the water was determined by comparing a comprehensive number of measured parameters to threshold values (Soltani et al. 2020; Hamlat et al. 2016). A powerful technique for combining all hydrochemical indicators into one value is the water quality index (WQI) (Soltani et al. 2021; Rachedi and Amarchi 2015). Horton (1965) proposed the concept of WQI to assess the quality of river water in the USA (Soltani et al. 2021). Many WQIs with particular specifications have been developed as a result of this process. Traditional indices suffer from fundamental drawbacks, including shortages of selection parameters—the majority of which do not account for toxic substances (heavy metals), a failure to take into account uncertainty, and subjective, deterministic formulations of the equation indices (Oukil et al. 2021; Ocampo-Duque et al. 2006). The methods used for classification evaluation of water quality, are matter element model (Wang et al. 2019), fuzzy synthetic evaluation (Zou et al. 2006), gray analysis method (Zhang et al. 2018), logistic curve model (Jin et al. 2003), attribute recognition model (Wang and Zou 2008), fuzzy logic (Yan et al. 2010), and k-nearest neighbors method (k-NN) (Modaresi and Araghinejad 2014). These methods necessitate data analysis expertise as well as understanding of water quality parameters. His methods are becoming more and more useful and well-liked for water quality issues as these limitations can be solved utilizing machine learning methodologies, making water quality monitoring based on sensor-generated data viable and affordable (Liao et al. 2011). However, some conventional methods for water quality assessment are unsuitable and, as a result, unable to provide a better performance for real-time applications due to high computational time and complexity, as well as proper inaptitude and incapacity due to nonlinear complicated relationships between all monitoring parameters and qualitative status. In the last few decades, robust methodologies and indices to classify water quality status have been developed. These approaches and indices solve previous limitations through natural language reasoning and successful approximation of the calculated index among complicated combined parameters. Neural network algorithms are frequently considered a solution to this kind of modeling process among others (Areerachakul and Sanguansintukul 2010). Methods like support vector machines (SVMs) and artificial neural networks (ANNs) were efficiently used in several applications (Liao et al. 2011; Wu et al. 2007). In this study, the monitoring process is a multiclass classification problem, but the development of WQI is a regression problem that transforms the several parameters containing water into one single number to describe the allover water quality. Many studies have employed several water quality indicators (WQIs) to evaluate the water's suitability for human consumption utilizing a variety of factors that must be carefully selected in order to get significant results (Oukil et al. 2021; Abbasi and Abbasi 2012). The water quality index number may not accurately reflect the current state of the water's quality because even one poor parameter value may change the water quality index's entire story. Since it reflects general water quality, it does not represent any particular usage of water (Phadatare and Gawande 2016). Several parameters reflect an economic impact on the overall cost of the control and monitoring system (reduced physical sensors). The number of studies applying ANN- and SVM-based models that have been extensively employed in water quality monitoring has considerably increased in these recent years (Liao et al. 2011; Phadatare and Gawande 2016). SVMs, a class of data-based learning algorithms that are relatively new and were first described by Vapnik (1995) (Vapnik 2000), have come to be used as an alternative approach in hydrologic research fields where ANNs are most commonly used. Most SVM applications have been focused on surface water problems. Yoon et al. (2011) applied ANN and SVM in their case studies. They concluded that the SVM model performance was better than ANN. The traditional neural networks are greatly reliant on datasets and problems of the local optimum in the training phase, resulting in bad learning results of the model and the limitations of classic artificial neural networks resulting in no memory being associated with the model, which is a problem for sequential data, like text or time series? The statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of SVMs. The SVM method is considered one of the strong and universal classifiers and approximators with a highly desired degree of accuracy in machine learning, solving the problem of gradient disappearance in traditional ANN (Nieto et al. 2015).

To be efficient, the preparation of datasets requires a certain treatment, which ensures that the classifier models are well decided. Recently, the use of features selection for dataset treatment in classification applications has received significant attention which can increase the performance of classifier (Widodo and Yang 2007; Cao et al 2003). However, we need to select features to prevent the curse of dimensionality phenomenon and redundancy, since irrelevant features would decrease the classifier's accuracy (Widodo and Yang 2007).

The curse of dimensionality with high-dimensional data refers to the phenomena that occur when classifying data in order to train a precise model. Therefore, the selection of a subset of relevant and useful features without any transformation is desirable (Yang et al. 2006). Many approaches to feature selection are based on linear methods such as principal component analysis (PCA) (Widodo and Yang 2007). The use of feature selection can improve the performance of classification results by reducing the number of data inputs needed to attain training with a significant database and reducing running time (Widodo and Yang 2007; Kumar et al. 2005). The purpose of this paper is to integrate PCA in combination with SVM and ANN in the evaluation of the quality of water. The PCA is employed precisely here as a features selection method to easily describe the correlations between a list of variables (Ladjal et al. 2020) in the best way, by generating a set of orthogonal principal components, i.e., not correlated, thereby reducing the dimensionality of the original dataset. SVM and ANN are employed as multiclass classifiers based on three classes of water quality. A comparative study is examined with and without the features selection process. The combination data fusion by decision templates of two classifiers is performed. To the best of our knowledge, the classification of water quality status via data fusion that use the decision templates method has not yet been performed, and there are no references in this application that make use of the suggested methodology. Decision templates combine seamlessly the outputs of the best models of both SVM and ANN classifiers to enhance accuracy of the water quality classification.

In order to develop better classification tasks with ANN, SVM has a good potential for achieving effective data representation. The performance of the classification task is greatly enhanced by PCA, which helps in reconstructing the input representation and converts it to a reduced feature representation of data related to the input data. The main contributions of this work are as follows:

1. We used a novel approach based on the SVM framework with ANN, studying the potential of our proposed approach to achieve an effective representation and dimensionality reduction using PCA method for the improvement of the binary classification results of shallow and traditional supervised machine learning algorithms.
2. PCA reduces a set of features that may be correlated into a smaller set of uncorrelated features or variables, called as PCs, which reflect an impact on the control and monitoring system's overall cost (reduced physical sensors). The selected parameters represent the overall quality of the water.
3. The combination data fusion using decision templates of two classifiers are applied in water quality monitoring. Decision templates combine seamlessly the outputs of the best models of both SVM and ANN classifiers to enhance the accuracy of the water quality classification
4. When compared to the results of similar approaches, better or at least equal and competitive results are obtained. Additionally, our method significantly actually reduces on training and testing times.

This study is particularly limited to experimental work carried out using datasets collected from the study area. The choice of the suitable hybrid approach is the object of this work by the application of SVM and ANN multi-class methods combined with PCA and using decision template's rule combination data fusion with respect to recognition rates, training times, and sensitivity to the noise.

Related works

Various methods, including support vector machines and artificial neural networks (ANN), have been employed in the selection of features and classification of water quality (SVM). The problem of classification with feature selection is crucial to data mining and machine learning. It has been used in a variety of applications in the real world. A user must first collect a set of training samples that are labeled with specified classifications in order to develop a classifier. A classification algorithm is then applied to the training of selected data to build a classifier that is subsequently employed to assign the predefined classes to test instances (for evaluation) or future instances (for application). Haghiabi et al. (2018) investigated the performance of artificial intelligence techniques that include the artificial neural network (ANN), the group data management method (GMDH), and the support vector machine (SVM) to predict the components of the water quality. During the development process of ANN and SVM, it was found that tansig and RBF as transfer and core functions have the best performance among the tested functions. Chou et al. (2018) applied ANN, SVM, regression trees, and linear regression to determine the water quality in the reservoir using data collected over 10 years in Taiwan. The ANN model was more accurate than the other unique models, sets, and metaheuristic regression hybrids

(Gakii and Jepkoech 2019). Mohammadpour et al. (2014) and Muharemi et al. (2018) employed SVM and ANN to water quality data. The SVM algorithm is competitive with neural networks (Gakii and Jepkoech 2019; Mohammadpour et al. 2014). The best result was achieved by using the artificial neural network with non-linear autoregressive (Muharemi et al. 2018).

Feature selection is one of the crucial steps for a comprehensive classifier. Many approaches, such as principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA), were used for feature selection and data reduction to improve the accuracy of the classification analysis. Because of the fact that a small set of uncorrelated variables is much easier to understand and use in further analysis than a larger set of correlated variables, this data compression technique has been widely applied to virtually every substantive area. Dilmi et al. (Dilmi and Ladjal 2021) employed LSTM RNNs and SVM with the three feature selection techniques. Additionally, we used three methods of cross-validation and two methods of the out-of-sample test to estimate the performance of LSTM RNN model. From the results, we found that the integration of LSTM RNNs with LDA and LSTM RNNs with ICA yields an accuracy of 99.72%, using Random-Holdout technique (Dilmi and Ladjal 2021).

Soltani et al. (2021) developed a new water quality index (WQI) based on Data Envelopment Analysis (DEA) to assess the water quality of 47 dams in Algeria. The development of the WQI in this kind of research is a regression problem that combines the several water-related data into a single value to represent the overall water quality. This novel approach has demonstrated its efficacy not just for classifying or evaluating areas based on water quality but also as an alternate tool to help decision-makers with resource management and funding allocation (Soltani et al. 2021). Also, Soltani et al. (2020) employed several techniques under the same framework, including the Canadian Council Ministers Environment Water Quality Index (CCME-WQI), principal component analysis and factor analysis (PCA/FA), the K-means clustering, and the ordinary least square (OLS) analysis (Soltani et al. 2020). Oukil et al. (2021) introduced a new approach, based on a unified framework incorporating data envelopment analysis (DEA) and ordered weighted averaging (OWA), for assessing water quality in contextual settings that involve a large number of hydrochemical parameters by WQI development. Instead of using pre-established borders, the k-means analysis was used to group the water quality of the wells into excellent, good, permissible, and unsuitable. Only one water source has been identified as excellent, whereas 17.65%, 45.10%, and 35.29% of the sampled wells, respectively, are categorized as good, permissible, and unsuitable water quality (Oukil et al. 2021). Chen et al. (2020) used large data with various parameters to examine the performance of the water quality prediction from the main rivers and lakes in China by applying 10 learning models (7 traditional and 3 ensemble models). The outcomes demonstrated that learning models could perform better in the prediction of water quality with larger datasets.

Multisource data fusion was used by Jiang et al. (2021) to combine a deep learning method with a linear method (multiple linear regression, MLR), as well as a more conventional learning algorithm (multilayer perception, MLP). These approaches take some indicators into account to comprehensively analyze and predict the drainage water quality in a city in southern China. The results showed that the deep learning algorithm, which consists of recurrent neural networks (RNNs), long-short term memories (LSTM), and gated recurrent units (GRUs), has good predictive performance, with GRU showing superior ability in predicting the chemical index of water quality and a faster learning curve (Jiang et al. 2021). Where the input data contains sequences that are too long, RNNs and LSTMs are quite good at extracting patterns in the input feature space. While traditional linear models can be difficult to adapt to problems with multiple or many inputs, they can nearly accurately depict problems with problems with many input variables, which is especially valuable in forecasting time series (Gakii and Jepkoech 2019).

To our best knowledge, the previous works were performed using single and separate models. In this work, we have presented a specific classification approach based on output data combination fusion to suit water quality monitoring efficiently.

Proposed framework

In this study, water quality assessment can be examined as a multiclass classification modeling problem. In general, it includes data acquisition and processing, selection of features, and classification of water quality. Figure 1 illustrates our proposed framework, based first on preparing datasets using PCA algorithms for feature selection before entering into ANN and SVM classifiers. Decisions concerning water quality status are obtained by using only a subset of appropriate and usable characteristics, without any transformation. In order to enhance the recognition rates, decision template combination data fusion is performed.

The aim is to classify the quality of water into three separate classes from the selected variables of water quality (I: excellent, II: middle, III: mediocre) according to the local environmental water quality guidelines. The next sections contain a short description of the principal methods employed in this paper.

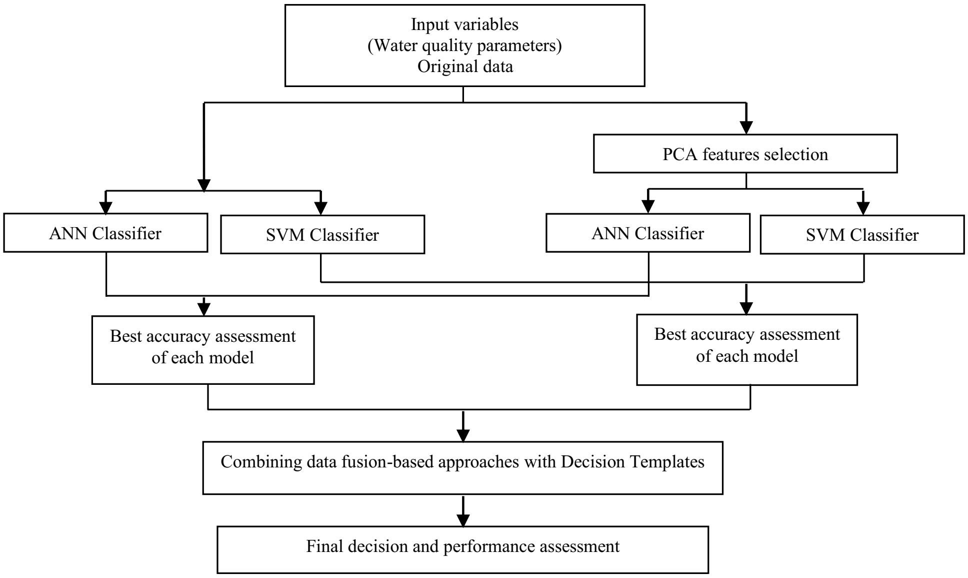
Input variables
(Water quality parameters)
Original data
PCA features selection
ANN Classifier
SVM Classifier
ANN Classifier
SVM Classifier
Best accuracy assessment of each model
Best accuracy assessment of each model
Combining data fusion-based approaches with Decision Templates
Final decision and performance assessment

Fig. 1 Flowchart of combining data fusion–based approaches with decision templates

Principal component analysis based on features selection

PCA is a technique commonly used to reduce multivariate problem dimensions. It was used in many areas like feature extraction and selections, high-dimensional data visualization, cluster analysis, pattern recognition, classification, and regression. Principal components (PCs) have a minimum loss of input data information in this method.

Without any transformation, PCA transforms a subset of features that may be correlated into smaller uncorrelated features or variables called PCs (Areerachakul and Sanguansintukul 2010). All of these factors are orthogonal to each other, so no redundant features exist. By the following equation, PCs can be defined (Jolliffe 2002):

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \dots + a_{im}x_{mj} \tag{1}$$

where z_{ij} represents PCs, a_{im} the related eigenvectors, and x_{mj} input features; i is the component number, j is the sample number, and m is the total number of features (Ladjal et al. 2020). This information is obtained through the resolution of the equation (Semmlow 2004):

$$|R - I\lambda| = 0 \tag{2}$$

where I is the unit matrix, λ is the eigenvector, and R is the variance–covariance matrix.

PCA results are generally evaluated by means of component values (values of the transformed features that match a certain data point) or loading values, also known as factor scores (weight to multiply each standardized original function in order to achieve the component score). We have N samples of M-dimensional data:

Step 1: To implement PCA, we should first calculate the variance–covariance matrix.
Step 2: Search for the matrix of the eigenvectors and diagonal matrix components as variance–covariance matrix values.
Step 3: Sort the PC eigenvectors in the decreasing order of importance of eigenvalues.
Step 4: By taking the dot product between the determined data and eigenvectors, project the data input into the directions of sorted eigenvectors.
Step 5: Based on the containment of a specified percentage of variability, select the first few PCs.

The PCs have the following characteristics (Jolliffe 2002):

- They are uncorrelated.
- They have sequentially maximum variance.
- The mean-squared approximation error in the representation by the first, several PCs of the initial inputs are minimal.

Artificial neural network

As shown in Fig. 2, artificial neural networks (ANN) are multi-layer, completely connected neural networks. ANN architectures have been classified into different types based on their training mechanisms and other features. By using the multi-layer perceptron (MLP) architecture (Jin et al. 2003), non-linear data classification is carried out. Many recent studies have used effective ANN models for water

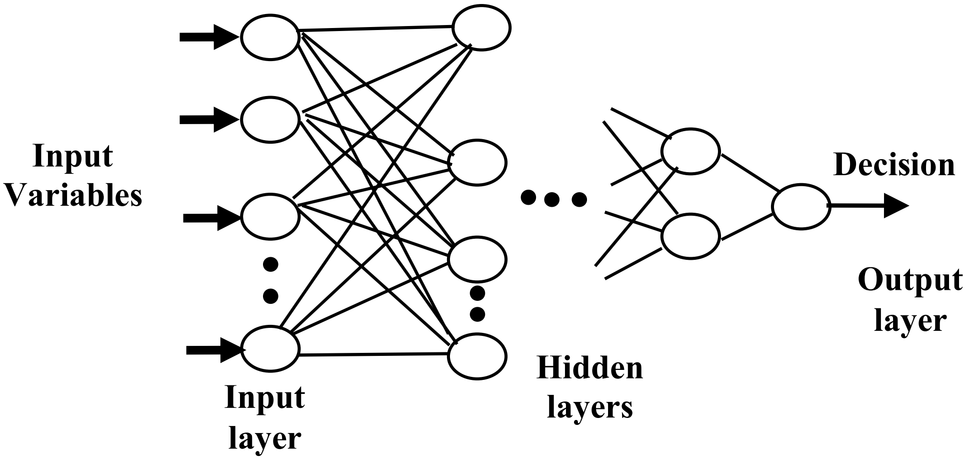

Fig. 2 Example of a multi-layer perceptron

quality monitoring (Zou et al. 2006; Liu and Zou 2012). The supervised learning process consists of calculating the weights that reduce the differences in the training set between the target output values y_r and the computed output values y_d.

The mathematical expressions of the *hyperbolic tangent* activation function and a minimum of the problem of quadratic optimization are, respectively, described by

$$f(u) = \tanh(\beta u) = \frac{e^{\beta u} - e^{-\beta u}}{e^{\beta u} + e^{-\beta u}} \tag{3}$$

$$C_w = \frac{1}{N}\sum_{i=1}^{N}(y_{r_i} - y_{d_i})^{2} \tag{4}$$

N presents the sample number of the learning dataset.

We used in this application the algorithm of Levenberg–Marquardt that is known as the 2nd order method, and it is rather better because they supply however much good results.

Support vector machines

The SVM method developed by Vapnik has been extensively used for classification, regression, and density estimation (Vapnik 2000; Schölkopf et al. 2002). In this method, through the construction of the optimal hyperplane, which is evaluated to optimize the generalization potential of the classifier, an initial input data space is mapped in a higher dimension space by selecting some non-linear functions, called kernel functions (Übeyli 2009).

Non-linear SVM classification

The initial input data should be implicitly mapped to a typically higher-dimensional feature space using kernel methods. In this mapping space, the classification method is then performed via the construction of the optimal linear separating hyperplane (Fig. 3). In this mapping space, the classification process is then carried out by building the optimal linear separating hyperplane (Fig. 3) with a maximization margin between it and the nearest point to obtain high generalization capacity via the quadratic optimization problem. The problem of quadratic SVM optimization for binary classification was established by the following dataset:

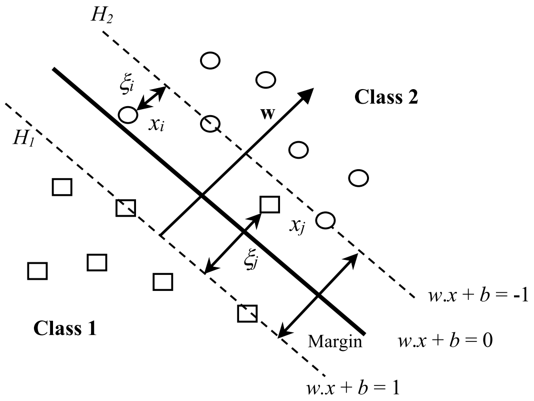

Fig. 3 The optimal hyperplane and margin of a binary SVM

$$(x_i, y_i), y_i \in \{-1, +1\},\ i = 1, \ldots, n \tag{5}$$

n is the number of observations, and $x \in \Re^d$ and y_i is a distribution and the corresponding class label respectively.

The optimal separating hyperplane is determined by the vector of weight w and a constant b, defined by Ladjal et al. (2020) and Bae et al. (2010):

$$w.x + b = 0 \tag{6}$$

Under constraints

$$\begin{aligned} &\left(w.x_i\right)+b \geq +1,\ \mathit{if}\ y_i=+1 \\ &\left(w.x_i\right)+b \leq -1,\ \mathit{if}\ y_i=-1 \end{aligned} \tag{7}$$

The primal quadratic minimization problem according to w, b is given by (Singh et al. 2011)

$$\begin{cases} \min_w\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \mathit{with}\ \ y_i(wx+b) \geq 1-\xi_i\ \ \xi_i \geq 0,\ i=1,\dots,n \end{cases} \tag{8}$$

The dual quadratic maximization problem using multipliers of Lagrange α_i is given by (Singh et al. 2011; Hend et al. 2010)

$$\begin{cases} \underset{\alpha_i}{Max}\ \ L(\alpha)=\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \\ \mathit{with}\ \sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leq\ \alpha_i \leq C\, i=1,\dots,n \end{cases} \tag{9}$$

C is a penalization parameter that controls the level of errors in classification.

The nonlinear mapping $\prec \Phi \succ$ is carried out via a kernel function $K(x_i, x_j)$ from an input space to some higher dimensional feature space (Ladjal et al. 2020). The new dual quadratic maximization problem is defined by (Hend et al. 2010; Horng 2009)

$$\begin{cases} \underset{\alpha_i}{\max}\ L(\alpha)=\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \\ \mathit{with}\ \sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C \end{cases} \tag{10}$$

Thus, the vector solution $\alpha^0 = (\alpha_i^0, \dots, \alpha_n^0)$. Using the theorem of *Karush–Kuhn–Tucker (KKT*) for an optimal weight vector, α is (Horng 2009)

$$\alpha_i^0\left\{y_i[(w_0 x_i)+b_0]-1\right\}=0, i=1,\dots,n \tag{11}$$

This means $\alpha_i^0 = 0$ or $y_i[(w_0 x_i)+b_0] = 1$, the letter corresponds to the support vectors (SVs) present at the nearest point to optimal hyperplane, which is equivalent to (Horng 2009)

$$SVs = \left\{x_i \text{ that } \alpha_i \succ 0\right\} \tag{12}$$

The function of decision is defined by (Singh et al. 2011; Horng 2009)

$$f(x) = sign(\sum_{SVs}\alpha_i y_i K(x_i.x)+b) \tag{13}$$

If $f(x)<0$, then x belongs to class -1; if not, it belongs to class 1, as b is the solution of the equation (Ladjal et al. 2020).

We can use all functions that satisfy *Mercer*'s theorem as a kernel function with appropriate parameter selection for more performances. The widely employed kernel functions are (Vapnik 2000; Schölkopf et al. 2002; Abedi et al. 2012):

The polynomial function:

$$k(x,x') = (\gamma.x^T.x' + c)^d \tag{14}$$

with $c \geq 0$ and $d \in N$

The radial basis function (RBF):

$$k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \tag{15}$$

One-against-all approach

The "one-against-all" approach is commonly used in multiclass classification problems (Burges 1998; Deng et al. 2011). Consider a multi-class k-class problem, where we can have N examples of training set:$\left\{(x_1,y_1),\dots,(x_N,y_N)\right\}$. Here, $x_i \in \mathfrak{R}^k$ is a k-dimensional input feature vector, and $y_i \in \{1, 2, \dots, k\}$ is the corresponding class output. The one-against-all approach (OAA) constructs k binary SVM models where the number of classes needed is k. With all the training samples in the ith class with positive class and all the other samples with negative class, the ith SVM is trained. The ith SVM defines the following optimization issue that results in the final decision function: $f_i(x) = w_i^T\varphi(x) + b_i$ (Wu et al. 2007; Deng et al. 2011):

$$\begin{aligned} &\underset{w^i,b^i,\xi^i}{\text{minimise}}\ \frac{1}{2}\|w^i\|^2 + C\sum_{i=1}^{N}\xi_j^i(w^i)^T \\ &\text{subject to } \left({}^{w^i}\right)^T\varphi(x_j)+b^i \geq 1-\xi_j^i,\ \ \text{if}\ \ y_j=i, \\ &\qquad\qquad \left({}^{w^i}\right)^T\varphi(x_j)+b^i \leq -1+\xi_j^i,\ \text{if}\ \ y_j \neq i, \\ &\qquad\qquad \xi_j^i \geq 0,\ j=1,\dots,\ N, \end{aligned} \tag{16}$$

where $\widehat{y}_j = 1$ if $y_j = i$ and $\widehat{y}_j = -1$ otherwise.

Sample x is classified as in class i^* at the classification step, whose f_{i^*} produces the largest value (Wang and Yang 2010):

$$i^* = \arg\ \max_{i=1,\dots,k} f_i\ (x) = \arg\ \max_{i=1,\dots,k}\ (w_i^T\varphi(x)+b_i). \tag{17}$$

Methods of classifier combination fusion

Integrating information from multiple sources and making combined decisions from them is becoming a common task across several disciplines and applications. A simple set of well-known combination data fusion methods such as minimum, maximum, majority voting and average compared with decision templates has been broadly applied to build a multiple classifier model for our proposed approach.

Conventional fusion methods

Average, maximum, and minimum There is a similar idea to these approaches. The maximum method selects the largest value for each class among the outputs of the classifiers. Then, the limit is compared, and a class with a greater value is chosen. It is calculated as follows for a multi-class problem (M) with L classifier models (Kuncheva et al. 2001):

$$\max_{z=1,\ldots,M} = \left\{ \max_{y=1,\ldots,L} \left\{ d_{y,z}(x) \right\} \right\} \tag{18}$$

Here, $dy,z(xi)$ is the degree of support determined, by the y^{th} classifier for the example x of the class z. The average and the minimum methods are the same as the maximum method except that the smallest values are compared as (Kuncheva et al. 2001; Min and Cho 2007)

$$\max_{z=1,\ldots,M} = \left\{ \min_{y=1,\ldots,L} \left\{ d_{y,z}(x) \right\} \right\} \tag{19}$$

For the average, the maximum methods compare the mean values (Kuncheva et al. 2001):

$$\max_{z=1,\ldots,M} = \left\{ \underset{y}{avg} \left\{ d_{y,z}(x) \right\} \right\}, \underset{y}{avg} \left\{ d_{y,z}(x) \right\} = \frac{1}{L} \sum_{y=1}^{L} d_{y,z}(x) \tag{20}$$

Majority voting The theory of this system is that the votes obtained from each classifier are counted and that the class with the highest number of votes is affected (Ruta and Gabrys 2005).

Decision Templates

The application of decision templates (DT) as a method of classifier combination fusion was proposed by Kuncheva (Kuncheva et al. 2001). DT is a method which makes employs all the base classifiers used on each of the m templates (or m datasets — one per class) with the same training set that is used for the set of classifiers (Haghighi et al. 2011). For the m multi-class problem, the classifier decisions can be organized in an output profile (DP(x)) as a matrix. The DP(x) for example x is a matrix composed of the $d_{t,j}$ $\epsilon[0, 1]$ elements representing the support defined by the tth classifier to class φ_j. Decision templates DT_j are the averaged output decision profiles obtained from X_j, the set of training examples belonging to the class φ_j (Zhang et al. 2014; Chen et al. 2010):

$$DP(x_i) = \begin{bmatrix} d_{1,1}(x_i) \ldots. & & d_{1,M}(x_i) \\ & d_{y,z}(x_i) & \\ d_{L,1}(x_i) & & d_{L,M}(x_i) \end{bmatrix} \tag{21}$$

$$DT_j = \frac{1}{\left| X_j \right|} \sum_{x \in X_j} DP(x) \tag{22}$$

where $dy,z(x_i)$ is the degree of support defined by the y^{th} classifier for the example x_i of the class z, and L is the number of classifiers in an ensemble. When decision output profiles are generated, the template of the class m is predicted as follows (Min and Cho 2007; Zhang et al. 2014):

$$DT_m = \begin{bmatrix} dt_m(1,1) \ldots. & & dt_m(1,M) \\ & dt_m(y,z) & \\ dt_m(L,1) & & dt_m(L,M) \end{bmatrix} \tag{23}$$

$$dt_m(y,z) = \sum_{l=1}^{n} u_{m,l} d_{y,z}(x_l) / \sum_{l=1}^{n} u_{m,l} \tag{24}$$

The similarity S between the decision template DTj for a class φ_j and the decision output profile for a defined test example x is

$$S_j(x) = 1 - \frac{1}{T \times C} \sum_{t=1}^{T} \sum_{k=1}^{C} \left[D_j(t,k) - d_{t,k}(x) \right]^2 \tag{25}$$

The last final decision of the ensemble is determined by assigning the test example to the class with the biggest similarity:

$$D(x) = \text{argmax}_j S_j(x) \tag{26}$$

The similarity between the decision output profile of a test example and each prototype is identified in the test process. In the class of the most comparable prototype, the example is then affected. Kuncheva (Kuncheva et al. 2001) studied DT with various distance measurements and achieved great success in classification compared to traditional combination data fusion techniques (Min and Cho 2007).

Results and discussion

In this study, the aforementioned proposed framework was applied to water quality data from Tilesdit station in Bouira (Algeria). For testing the applicability of the suggested methodology, our monitoring model consists of three steps: features selection and recognition of the water quality status with data combination fusion. The feature selection technique is based on PCA, and the classification technique is based on SVM and ANN multi-class methods combined using decision template's rule combination data fusion. The hardware used to perform our simulation experiments are as follows: we have used an Intel Core TM i7-6820HQ and 2.71 GHz CPU processor with 8 GB of memory. All proposed methods were implemented and assessed using

MATLAB2019b environment software with Windows 10 (64-bit) operating system.

Study area and data descriptions

The study area (Tilesdit dam, Fig. 4), is situated in the region of Bechloul, 20 km southeast of Algeria's Bouira Department (Ladjal et al. 2016). It is located approximately 122 km east of Algiers (35° 13′ 22″ N 4° 14′ 23″ E) (Fig. 4). The research area is characterized by a semi-arid climate. Mean annual temperatures range from 20.4 to 37.9 °C. Yearly precipitation averages are about 440–660 mm/year.

The volume of reservoir was evaluated in March 2007 at 167 million m^3. Water from the dam needs to be designed to curb the tension in water distribution in 12 cities. The transfer of water including the launch of the construction works was scheduled for early 2011. Work is underway to connect many towns of the Bouira Department. A processing plant with a capacity of 74,000 m^3/day is equipped. Water collected in the dam is pumped to the treatment plant. This being at the same pace is commissioned since 2009. It performs the purification process through the five processing levels: pre-treatment, pre-oxidation, clarification, disinfection, and refining. The clarification step is performed by the method of coagulation-flocculation, decantation, and filtration through a phase separator and a sand filtration stage.

In this paper, we search to develop our framework approach of control and monitoring of water quality using several descriptors provided in a water production plant by certain physical sensors. These parameters are collected during 3 years from the Tilesdit production plant (2009–2018). The parameters like pH, temperature (T°), electrical conductivity (EC), and turbidity (TU) are collected by sensors installed in all treatment process of the station (Ladjal et al. 2020, 2016). Every week in the lab, some chemical parameters are examined such as magnesium (Mg), bicarbonate (B), total hardness (TH), and full title alkaline (FTA). The above-mentioned collected data will be applied to check the water quality assessment model. A summary table of statistical characteristics of the collected parameters of water under study is given by Table 1.

Table 1 Statistical characteristics of the collected parameters

Variables	Min	Max	Average	Standard deviation
pH	7.15	8.30	7.57	0.25
EC (ms/cm)	414.00	624.00	585.40	36.28
T (°C)	9.70	24.20	16.13	3.48
TU (NTU)	1.32	23.81	3.,83	2.39
Mg (mg/l)	7.29	47.63	22.27	4.93
B (mg/l)	158.62	289.14	222.50	23.21
TH (mg/l $CaCO_3$)	0.00	168.00	32.29	23.03
FTA (mg/l $CaCO_3$)	130.00	237.00	181.84	18.70

Data features selection

For data features selection, the PCA method is used with 80–90% variation of eigenvalues, without any transformation of the resulting components which are uncorrelated (De León 2006). A total of 1800 samples from eight physicochemical water quality parameters are used in this phase (Fig. 5).

A variance–covariance matrix is formed by using PCA on input variables. Eight eigenvalues are obtained after solving Eq. (2). Table 2 presents the PCA results and statistical parameters such as eigenvalues, cumulative variance proportion, and variance proportion. The four PCs indicate 84.68% of the total input samples variance proportion and eliminate the remaining components, as set out in Table 3. These PCs calculate mainly the initial data variance.

In addition, PCA applications are used to obtain eigenvectors to evaluate the coefficients for the training of PCs. The correlations between each variable and the main acquired components are shown in Table 2. In this table, the most effective parameters in PCs training are exposed in bold

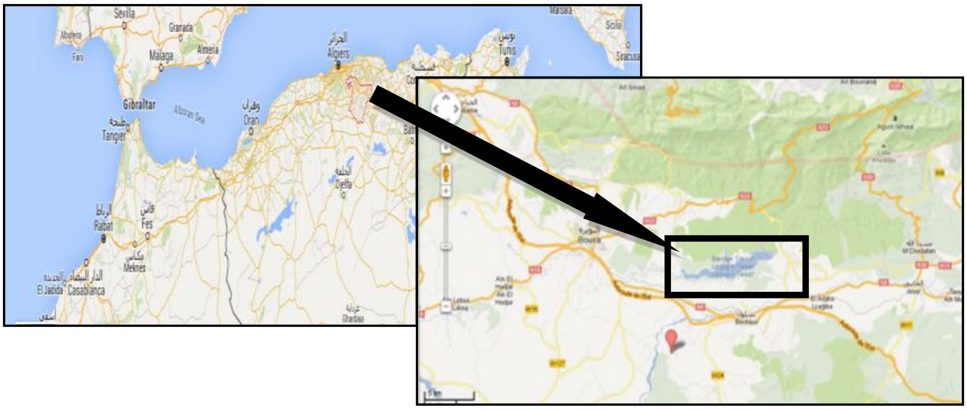

Fig. 4 Map showing the Tilesdit dam–Bouira–Algeria (Google Maps)

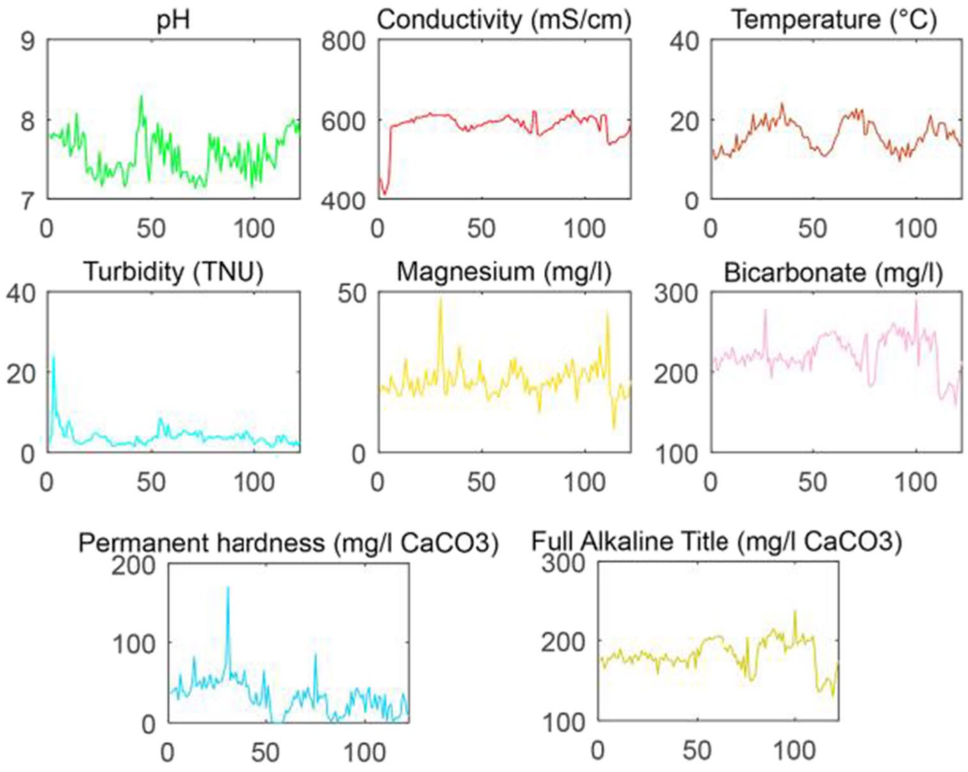

Fig. 5 Evolution section of the water quality variables [4]

Table 2 Statistical characteristics of the resulted PCs

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Eigenvalues								
	2.57	2.23	1.12	0.85	0.57	0.33	0.27	0.06
Total variance proportion (%)								
	32.07	27.93	14.05	10.63	7.16	4.10	3.32	0.74
Cumulative variance proportion (%)								
	32.07	60.00	74.05	**84.68**	91.84	95.93	99.26	100
Variables of eigenvectors obtained by applying PCA								
pH	−0.57	−0.46	−0.25	**0.52**	−0.08	0.29	0.20	0.00
EC	**0.74**	0.30	−0.37	0.08	0.26	0.33	−0.21	0.00
T°	−0.04	**0.78**	−0.10	−0.45	−0.28	0.22	0.22	0.00
TU	−0.26	−0.47	**0.73**	−0.27	0.07	0.29	−0.10	0.00
Mg	0.38	0.48	0.43	0.48	−0.44	0.03	−0.13	−0.01
B	**0.87**	−0.41	0.11	−0.03	0.01	0.01	0.20	−0.17
TH	−0.03	**0.70**	0.42	0.27	0.46	−0.02	0.20	0.02
FTA	**0.85**	−0.45	0.09	−0.00	−0.07	−0.01	0.16	0.17

font. The total variance in the dataset accounts for 84.68% of the first four principal components together. The first component (PC1) is 32.07%, with 27.93% being the second component (PC2), 14.05% being the third component (PC3), and 10.63% of the total variance being the fourth component (PC4). In general, the *EC*, *B*, and *FTA* concentrations are obvious to be the most effective for PC1 and represent more than 32% of input variable variance proportions. Furthermore, the *T*° and *TH* concentrations also have the most effect on the PC2, which contains more than 27% of input variables' variance proportions. Moreover, *TU* and *pH* concentrations are affected by PC3 and PC4 respectively (Ladjal et al. 2020).

In Table 2, the rapid decay of eigenvalues is apparent. For the evaluation of prevailing physicochemical processes, the eigenvalues of the first fourth principal components (PC1-PC4) can be used (Bhardwaj et al. 2010; Ayeni 2013). The *EC*, *B*, and *FTA* concentrations are highly

Table 3 Classification results using ANN models

Number of hidden layers	Number of neurons in hidden layers	Recognition rates (%)			
		Without PCA features selection (8 variables)		With PCA features selection (4 variables)	
		Training	Testing	Training	Testing
1	(4)	**97.56%**	**86.63%**	99.83%	97.94%
1	(8)	100%	84.18%	97.83%	98.11%
2	(4–8)	100%	84.15%	98.92%	98.07%
2	(10–10)	98.78%	83.86%	99.83%	98.42%
3	(4-8-12)	**97.56%**	**86.61%**	**99.75%**	**99.13%**

positive (0.74–0.87), while the *Mg* concentration is low positive for the first component (0.38). *T* ° and *TH* have high positive loads in the PC2 (0.70–0.78), and the other concentrations show low positive loads (0.3–0.48). The concentrations of *TU* in the PC3 have high positive loadings (0.73), while concentrations of *FTA* have low positive loads (0.09). The *pH* concentrations for the PC4 show high positive charges (0.52), while the *Mg* displays moderate positive charges (0.48), and *TU* and *TH* show low positive charges (0.08–0.27) (Ladjal et al. 2020).

From Table 2, the first four PCs are the input features of the evaluated multi-class classifiers. Variables retained are pH, temperature (*T*°), electrical conductivity (EC), and turbidity (TU). As a result, monitoring must take place at the treatment plant and in a continuous way using selected parameters that are the most representative used due to strong existing correlations between all the parameters, as well as the most fundamental and easily measured by physical sensors in the monitoring water quality system. These results are equivalent to the results obtained in literature (Ladjal et al. 2020, 2016) with different database and period, which adopts the same selected variables, and these parameters were measured in the field, using the station's sensors. In any case, this solution is not final; a relearning system should probably be conducted periodically so that situations that might arise can be taken into account and continuously adjusted to change water quality.

Samples classification with SVM and ANN

Traditional methods, in most drinking water production units, are based on knowing the various parameters of the raw water through chemical analyses carried out in the laboratory. These methods require human inspection and are time-consuming. This approach, in addition to the disadvantage of having a relatively long delay time, does not allow fine monitoring of the evolution of the raw water quality. There is a necessity to look into the water standards before usage. Water quality evaluation was assessed by comparing a long list of measured parameters with water standards. While it can be hard to probe into current practices and evaluate the methodologies of individual sources of pollutants, the quality of a water body, to deem the purity of it, laboratory practices that are labor-intensive and time-consuming need an automated technological alternative. Automatic machine learning facilities supply machine learning with a push of a button or, on a minimum level, ensure retaining algorithm execution; data pipelines and code, generally, are kept from sight and are anticipated to be the steppingstone for normalizing AI. However, it is still a field under research. This is a problem that can greatly benefit from artificial intelligence (AI) like ANN and SVM methods which makes it a good water quality classification tool and serves as a basic tool for decision support. It is wise to assume that the water quality monitoring operation can be seen as a pattern recognition problem, where the pattern represents the measurements related to water parameters, and the outputs correspond to the different water statuses. The proposed model could be considered an effective tool for identifying the water quality status. In addition, the major advantage of the proposed model is that it could be useful for ungagged catchments or those lacking enough numbers of monitoring stations for water quality parameters.

To carry on the training and classification process, datasets for the training and test phases are developed and arranged in three separate classes of water quality (I: excellent, II: middle, III: mediocre) according to the local environmental water quality guidelines (Décret 2011). A collected data collection of 1800 samples (Table 1) was used.

In this work, diverse architectures of ANN and hyperbolic tangent activation functions have been applied to the hidden and output layers to establish the suitable number of hidden layers and neurons (Adem et al. 2019). The SVM using OAA approach is used to carry out the multiclass classification process. *C*, *d* and σ are the three parameters associated with the SVM kernel functions. The parameter *d* related to the polynomial degree and σ for RBF function, and *C* is the penalization factor. Therefore, good choice of all parameters in the two models ANN and SVM can show excellent results and in the opposite case can cause under fitting or over fitting problem (Widodo and Yang 2007).

To assess the two used methods, tenfold cross-validation has been performed in training and testing phases. The cross-validation process can stop the over fitting problem that is very important in subsample random selection used for testing and training datasets. We can develop classification models with high performance and accuracy through the use of cross-validation.

Table 3 indicates the results of ANN multi-class models using data input with and without features selection. The different parameters of training, such as the global number of neurons and hidden layers and the recognition rates (training and testing), evaluated all the samples by using the correct classification rate.

Table 4 indicates the results of sample classification with SVM multi-class models. The performance criteria, such as number of support vectors (*NSV*), and the recognition rates for training and testing phases are determined for different kernel functions and its parameters and values of factor *C*.

In Tables 3 and 4, the recognition rates on the original dataset without features selection process are more than 96% in training step and from 83.86 to 89.09% in the testing step for the two models (ANN and SVM). The existence of irrelevant and useless features decreases the performance of the classification process (Widodo and Yang 2007). Then, as shown in the cited tables, the recognition rate with PCA features selection ranged from 97.83 to 99.83% in the training step and from 97.80 to 99.13% in the testing step for the two models (ANN and SVM). It is better than the precedent classification without feature selection.

As shown in Table 4, the effect of choice of architecture and parameters of networks is important. Indeed, a good choice of the ANN architecture characteristics can improve the performance of classification. The best model for this application is the network with three hidden layers using the original feature set with and without the features selection process. This architecture is characterized by a recognition rate in the testing step with a features selection process of 99.13%. For the SVM model, the feature selection step increased the performance of the classification process. It can be compared with Tables 3 and 4 in the case of with and without features selection by PCA. In Table 4, the recognition rate in the training phase for linear kernel is usually lower than polynomial and Gaussian RBF kernel with and without features selection. Even though the degrees of polynomials are 2, 3, and 4 in the process with features selection. However, the recognition rate reaches 98.48% using the Gaussian RBF kernel ($\sigma = 0.1$, $C = 1000$) due to the good quality of data input after the feature selection process. This model is characterized by the recognition rates in training and testing steps which are 99% and 98.48% respectively. Gaussian RBF kernel has shown to be the best choice for this application (Bouamar and Ladjal 2012; Widodo and Yang 2007; Djerioui et al. 2018). Therefore, a good choice of kernel function and its parameters *C* and σ or *d* and γ can achieve the best performance in classification steps (Widodo and Yang 2007).

The recognition rates of each model with features selection process are high. The features selection process searches the uncorrelated components from the input data using PCA which is useful to increase the performance of classification. The use of *k*-fold cross-validation (CV) with the SVM technique is typically the most appropriate method. These results are equivalent to the results obtained by Dilmi and Ladjal (2021). Generally, as listed in Tables 4 and 5, the SVM models using the strategy of OAA are better than ANN models mainly when using the original data without a features selection process in high-dimensional data classification (Widodo and Yang 2007). Moreover, using kernel parameters selection will increase the performance of classification. However, the ANN models are better than SVM models mainly when using the original dataset with the features selection process as listed in the table against the recognition rate in the testing step with a small improvement.

Noise robustness of classifiers

It is important to evaluate the robustness of techniques used to see their impact in the system decision. Adversarial examples, generated by adding small but intentionally imperceptible perturbations to normal examples, can mislead classifiers to make incorrect decisions. The key is to compare and analyze the data paths of both the adversarial and normal examples. In this step, all input data values were standardized and normalized between 0 and 1 to avoid having more weight being assigned to features with larger values. Normalization is essential to get rid of biases in data for their accurate analysis.

In order to eliminate dimension differences, the following equation was used for data standardization and normalization, and then all input and output data were standardized and normalized to the range [0, 1] (Msiza et al. 2008; Liu et al. 2013):

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{27}$$

where x is the initial datasets, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values, respectively.

The results showed that normalized data is easier to process (Liao et al. 2011). To address this issue, we added artificially white noise to the initial unperturbed data inputs. We study the noise stability of such methods on unperturbed inputs and observe that internal activations of adversarial trained networks have a lower signal-to-noise ratio (SNR).

Table 4 Classification results of SVM models and selected kernel parameters

Kernel parameters	Recognition rates (%)					
	Without PCA features selection (8 variables)			With PCA features selection (4 variables)		
	Training	Testing	NSV	Training	Testing	NSV
Linear ($d=1$, $\gamma=1$, $C=1000$)	96.34%	89.09%	19	98.50%	98.02%	49
Polynomial 2 ($d=2$, $\gamma=1$, $C=1000$)	100%	84.42%	28	98.67%	98.00%	53
Polynomial 3 ($d=3$, $\gamma=1$, $C=100$)	100%	85.78%	26	98.42%	98.20%	**73**
Polynomial 4 ($d=4$, $\gamma=1$, $C=100$)	100%	85.24%	37	98.92%	97.80%	50
Gaussian RBF ($\sigma=0.1$, $C=1000$)	**100%**	**86.88%**	**42**	**99.00%**	**98.48%**	63
Gaussian RBF ($\sigma=1$, $C=1000$)	100%	85.82%	45	99.50%	98.04%	40
Gaussian RBF ($\sigma=2$, $C=1000$)	100%	84.14%	53	98.83%	98.07%	67

Table 5 Recognition rates according to the SNR of PCA-SVM and PCA-ANN multi-class models

Input variables	Models with features selection	SNR (dB)	Methods	Recognition rate according to the SNR	MSE_ANN
(4 variables) (EC, T°, TU, pH)	Without noise		ANN	96.33%	0.04
			SVM	97.24%	
		10	ANN	–	1.4
	ANN		SVM	87.80%	
	(4–8-12)	20	ANN	83.17%	0.15
			SVM	97.17%	
	SVM	40	ANN	95.83%	0.05
	Gaussian RBF ($\sigma = 2^{-1}$, $C = 1000$)		SVM	97.17%	
		60	ANN	96.33%	0.04
			SVM	97.24%	

SNR is calculated via recognition rate using five different levels of white noise: 10, 20, 40, and 60 dB. A quantitative evaluation and a case study were conducted to demonstrate the robustness of the noise of our methods. Table 5 presents the results associated with ANN and SVM multi-class models using reduced real data input with additive white noise. The various parameters, such as the recognition rates according to the SNR which are calculated by using the correct classification rate of real data and the minimum square error (MSE_ANN) for the ANN model, are presented.

The results showed that normalized data is easier to process, but the recognition rate has dropped. Compared to ANN, we found a remarkable resistance to the white noise of SVM multi-class models. When the SNR is equal to or greater than 20 dB, the test vectors using SVM are absolutely insensitive to the different noise levels applied. This explains why this approach enjoys exceptional immunity. For the ANN model, when the SNR decreases, we observe a strong deterioration of the recognition rate. In addition, for a ratio lower than or equal to 10 dB, there is a really clear immunity limitation. The MSE explains this situation. However, it has been noted that when this SNR is equal to or greater than 40 dB, the test vectors do not exhibit concrete degradation. We may assume that the ANN model offers an appropriate resistance an SNR above 40 dB. Ultimately, this model tends to be more noise sensitive and therefore less stable than the SVM model. Table 6 summarizes the characteristic results corresponding to the two models carried out on the suggested real data.

It appears that the two models perform good results on the decisional level with recognition rates of more than 98% in the training and testing phases with the features selection process. In the training step, the SVM model is rather better positioned on the computing time, which gives it the benefit of integration into a dynamic monitoring system. With SNR above or equal to 20 dB, the two multi-class models have slightly strong immunity. A classifier is robust if it is insensitive to outliers and noisy data. Thus, what is gained in robustness and lost in precision? It is therefore necessary to carefully choose these thresholds so as to make a good compromise between precision and robustness (Saint-Jean and Frélicot 2001). Furthermore, the ANN model suffers from a significant handicap related to its apparent noise sensitivity. The SVM model removes this limitation because of its excellent noise robustness. Finally, we can conclude that the classification method performed using the SVM technique on a real Tilesdit dam data provides the best performance and the acceptable solution combined with the PCA features selection strategy. These results are equivalent to the results obtained by Achmad et al. (Widodo and Yang 2007). This result is important because it reflected an economic impact

Table 6 Characteristics of ANN and SVM models

Input variables	Models with features selection		Characteristics		
		Training time (s)	Recognition rate (%)		Robustness
			Training	Testing	
(4 variables) (EC, T°, TU, pH)	ANN (4–8-12)	65.29	99.75%	99.13%	Good (96.33% at 60 dB)
	SVM Gaussian RBF ($\sigma = 0.1$, $C = 1000$)	1.52	99.00%	98.48%	Excellent (97.24% at 60 dB)

on the overall cost of the control and monitoring system (less training time and reduced physical sensors).

Classification using decision template rule combination

The results were combined after obtaining the classification accuracy of the classifiers. In this study, the fusion of primary classification results is carried out in comparison with a set of well-known combination methods such as majority voting, minimum, maximum, average, and Bayes, using the rules of the decision template for classifier combination. These techniques are noted as the best among all combination data fusion methods in pattern recognition (Polikar 2006; Kuncheva 2014). Because of this, it is important to find the viability of the combined structure preferred in this work. Table 7 illustrates the hybrid data fusion of the two ANN and SVM classifiers and the overall accuracy of the fusion approaches, and also of the suggested framework.

Referring to Table 7, it can be seen that the fusion of the classifiers proposed significantly increased the accuracy of the classification and enhanced the efficiency of the multi-class system proposed. These results are equivalent to the results obtained by Kuncheva (2014), Chen et al. (2010), and Bigdeli et al. (2015) in different databases and applications. These results also are equivalent to the results obtained by Ladjal et al. (2016) and which examine the data combination fusion using Dempster-Shafer Theory and appear considerably consistent with our findings in this study with the decision template method. By comparison of the results shown in this table, it can be found that the two models ANN and SVM have a good ability in water quality monitoring. The classifier combination data fusion is used to increase the classification accuracy and efficacy of the proposed process, meaning that the ANN and SVM results obtained have been combined. It can be shown that with the feature selection process, the classification precision increased by up to 98%. It can be shown that when we use the decision template technique, our approach offers greater classification precision, offering an increase in the recognition rate. The approach of fusing multi-classifiers with a decision template at the decision phase is more powerful compared to other approaches. This result denotes the high capability of using the classifier combination. Furthermore, it is useful for practical purposes, so that the proposed technique can be used efficiently for monitoring water quality. This means that the precision of water quality can be greatly enhanced by applying several classifiers. This growth is a strong explanation for the efficacy of the process of data fusion and the principle of the decision template. The results obtained underline the use of multiple sources of knowledge for accurate monitoring of water quality.

Conclusion

In this work, we have provided a performance assessment for intelligent water quality monitoring of ANN and SVM multi-class models. The research area is the Tilesdit dam in Algeria. An adequate intelligent procedure was proposed based on surface water physicochemical variables. It included PCA features selection, ANN, SVM, and data fusion method. We studied the theory and the practical implications of the proposed techniques and their adaption to the used classifiers models to help practitioners towards the implementation of the new tools on the decision-making front. These techniques have shown good results concerning accuracies. PCA was successfully applied to the feature selection process; however, we carried out this approach to exclude irrelevant and redundant features. Therefore, the use and implementation in the field of water quality control of these approaches are well justified. We have used a cross-validation procedure, particularly in this study that can prevent over fitting problems by selecting random subsamples used for training and testing datasets. We trained the ANN and SVM models onto the real data input without and with feature selection to show the importance of this process. Because these are among the major sensor systems from which information is derived, an evaluation of the performance of the ANN and SVM using real data from such sensor systems should have practical implications for water quality classification. With the use of PCA as a reducing technique of the input variables, the obtained results showed clearly excellent performances with a slight improvement in terms of recognition rates. Indeed, the use of PCA features to select the number of input parameters is decreased,

Table 7 Combination data fusion of the two classifiers ANN and SVM

	Models with PCA features selection (4 variables: EC, *T*°, TU, pH)						
	ANN (4–8-12)	SVM Gaussian RBF ($\sigma = 2^{-1}$, $C = 1000$)	Majority voting	Maximum	Minimum	Average	Decision template
Recognition rate (%) in testing steps	99.13%	98.48%	98.80%	99.13%	98.98%	99.20%	99.24%

indicating a small number of sensors. That means, the PCA technique reflected an economic impact on the overall cost of the monitoring system. With this reduction operation, we can say that the storage of data in memory can be considered advantageous for the enrichment of the database collected by the expert system. In general, since it depends on several climatic and geographical parameters, continuous enrichment of the database is necessary. The computational time assigned to the training dataset for the PCA-SVM model is extremely fast, which confers the advantage of integration in a dynamic multi-sensor monitoring system. The ANN model, however, suffers from a handicap related to its apparent noise sensitivity. However, due to its best robustness, this limitation is eliminated by the SVM model in particular. The principle of the optimization algorithm is another significant feature of SVM relative to ANN; the solution has a global optimum, eliminating the use of gradient-based search techniques that can cover a local optimum. However, there are no clear rules for fixing the number of neurons and hidden layers in the ANN technique, which is a major concern for obtaining an optimal architecture. The results obtained showed that using a one-against-all multi-class approach, SVM can achieve high performance in classification in terms of recognition rate, training time, and robustness. Overall, the present application demonstrates SVM's promising results. The use of multiple sources of knowledge is highly successful in enhancing classification accuracy. For this analysis, decision template (DT) has high potential. The precision of the system decision can be enhanced by using the decision template fusion method. With this method, the power of each classifier to achieve a more powerful classifier was combined. The decision template has shown more success than each classifier. Furthermore, a recognition rate of 99.24% was obtained. Hence, the use of various sensors and multiple classifiers and subsequently combining them is strongly recommended for classification in water quality monitoring. The use of the DT algorithm for final decision-making improves the system's efficiency and the accuracy of the approved classification process. When chemical parameters are unable to be continuously measured, the precision of the system decision can be increased by using new input parameters or soft sensors. It is also important to assess the robustness of the proposed solution concerning noise. It should be noted that the domain's sensitivity and unexpected threats need greater efforts to optimize the system's immunity and to make more changes to reduce the risks to public health.

Acknowledgements This work is supported by the General Directorate of Scientific Research and Technological Development, Ministry of Higher Education and Scientific Research of Algeria. The authors thank the editor and reviewer for many helpful and constructive suggestions and remarks about an earlier draft of this article which improved the paper quality considerably. The authors would like to thank the engineers from the Tilesdit dam direction for their support and for providing the facilities for this investigation and free access to databases and valuable guidance for the field sampling.

Author contribution All the authors contributed to the study conception and design through material preparation, data collection, and analysis. All the authors read and approved the final manuscript.

Mohamed Ladjal: conceptualization, methodology, software, formal analysis, investigation, resources, data curation and collection, writing — original draft, writing — review and editing, visualization.

Mohamed Bouamar: supervision, project administration, conceptualization, formal analysis, writing — review and editing, visualization.

Youcef Brik: conceptualization, software, formal analysis, investigation, visualization.

Mohamed Djerioui: software, formal analysis, investigation.

Data availability All data generated or analyzed during this study are included in this published article; they are available from the corresponding author on reasonable request. For the purposes of privacy, all used data are confidential and cannot be made available.

Declarations

Ethics approval This article does not contain any studies with any participants performed by any of the authors.

Informed consent to participate and publish None.

Conflict of interest The authors declare no competing interests.

References

Abbasi T, Abbasi SA (2012) Water quality indices. 1st Edition, Elsevier, Hardback ISBN 978-0-444-54304-2

Abedi M, Norouzi GH, Bahroudi A (2012) Support vector machine for multi-classification of mineral prospectivity areas. Comput Geosci 46:272–283

Adem K, Kiliçarslan S, Cömert O (2019) Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. Expert Syst Appl 115:557–564

Areerachakul S, Sanguansintukul S (2010) Classification and regression trees and MLP neural network to classify water quality of canals in Bangkok, Thailand. Int J Intell Comput Res 1(2):30–37

Ayeni O (2013) Interpretation of surface water quality using principal components analysis and cluster analysis. J Geogr Reg Plan 6(4):132–141

Bae MH, Wu T, Pan R (2010) Mix-ratio sampling : classifying multiclass imbalanced mouse brain images using support vector machine. Expert Syst Appl 37(7):4955–4965

Bhardwaj V, Singh DS, Singh AK (2010) Water quality of the Chhoti Gandak River using principal component analysis, Ganga Plain. India J Earth Syst Sci 119(1):117–127

Bigdeli B, Samadzadegan F, Reinartz P (2015) Fusion of hyperspectral and LIDAR data using decision template-based fuzzy multiple classifier system. Int J Appl Earth Obs Geoinf 38:309–320

Bouamar M, Ladjal M (2012) Performance evaluation of three pattern classification techniques used for water quality monitoring. Int J Comput Intell 11(02):1250013

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2:121–167

Cao L, Chua K, Chong W, Lee H, Gu Q (2003) A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing 55(1–2):321–336

Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Liu F, Zuo M, Zou X, Wang J, Zhang Y, Chen D, Chen X, Deng Y, Ren H (2020) Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res 171:115454

Chen W, Zhang SW, Cheng YM, Pan Q (2010) Prediction of protein–protein interaction types using the decision templates based on multiple classier fusion. Math Comput Model 52(11–12):2075–2084

Chou JS, Ho CC, Hoang HS (2018) Determining quality of water in reservoir using machine learning. Eco Inform 44:57–75

Deng S, Lin SY, Chang WL (2011) Application of multiclass support vector machines for fault diagnosis of field air defense gun. Expert Syst Appl 38(5):6007–6013

Décret exécutif N° 11–125 du 22 Mars (2011) Relatif à la qualité de l'eau de consommation humaine, Journal officiel de la Republique Algerienne N° 18

De León HRH (2006) Supervision et diagnostic des procédés de production d'eau potable (Doctoral dissertation, INSA de Toulouse)

Dilmi S, Ladjal M (2021) A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. Chemom Intell Lab Syst 214:104329

Djerioui M, Bouamar M, Ladjal M, Zerguine A (2018) Chlorine soft sensor based on extreme learning machine for water quality monitoring. Arab J Sci Eng 44(3):2033–2044

Gakii C, Jepkoech J (2019) Classification model for water quality analysis using decision tree. Eur J Comput Sci Inf Technol 7(3):1–8 (June 2019)

Hamlat A, Guidoum A, Koulala I (2016) Status and trends of water quality in the Tafna catchment : a comparative study using water quality indices. J Water Reuse Desalination 7(2):228–245

Haghiabi AH, Nasrolahi AH, Parsaie A (2018) Water quality prediction using machine learning methods. Water Qual Res J 53(1):3–13

Haghighi MS, Vahedian A, Yazdi HS (2011) Extended decision template presentation for combining classifiers. Expert Syst Appl 38(7):8414–8418

Hend S, Al-Khalifa A, Al-Ajlan A (2010) Automatic readability measurements of the arabic text: an exploratory study. Arab J Sci Eng 35(2C):103–124

Horng MH (2009) Multi-class support vector machine for classification of the ultrasonic images of supraspinatus. Expert Syst Appl 36(4):8124–8133

Horton RK (1965) An index number system for rating water quality. J Water Pollut Control Fed 37(3):300–306

Jiang Y, Li C, Sun L, Guo D, Zhang Y, Wang W (2021) A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. J Clean Prod 318:128533

Jin JL, Liu L, Ding J, Fu Q (2003) Logistic curve model of groundwater quality evaluation. Environ Pollut Cont 25(1):46–48

Jolliffe IT (2002) Principal component analysis, Springer Series in Statistics, 2nd edn. Springer

Kumar R, Jayaraman V, Kulkarni B (2005) An SVM classifier incorporating simultaneous noise reduction and feature selection : illustrative case examples. Pattern Recogn 38(1):41–49

Kuncheva LI, Bezdek JC, Duin RP (2001) Decision templates for multiple classifier fusion : an experimental comparison. Pattern Recogn 34(2):299–314

Kuncheva LI (2014) Combining pattern classifiers: methods and algorithms. John Wiley & Sons

Ladjal M, Ouali MA, Lass MD (2020) optimization of SVM parameters with hybrid PCA-PSO methods for water quality monitoring. In 2020 International Conference on Electrical Engineering (ICEE). IEEE, pp 1–6

Ladjal M, Bouamar M, Djerioui M, Brik Y (2016) Performance evaluation of ANN and SVM multiclass models for intelligent water quality classification using Dempster-Shafer Theory. In: 2016 International Conference on Electrical and Information Technologies (ICEIT). IEEE, pp 191–196

Liao Y, Xu J, Wang W (2011) A method of water quality assessment based on biomonitoring and multiclass support vector machine. Procedia Environ Sci 10:451–457

Liu D, Zou Z (2012) Water quality evaluation based on improved fuzzy matter-element method. J Environ Sci 24(7):1210–1216

Liu S, Tai H, Ding Q, Li D, Xu L, Wei Y (2013) A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math Comput Model 58(3–4):458–465

Min JK, Cho SB (2007) Multiple classifier fusion using k-nearest localized templates. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, pp 447–456

Modaresi F, Araghinejad S (2014) A comparative assessment of support vector machines, probabilistic neural networks, and k-nearest neighbor algorithms for water quality classification. Water Resour Manage 28(12):4095–4111

Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA, Chan NW (2014) Prediction of water quality index in constructed wetlands using support vector machine. Environ Sci Pollut Res 22(8):6208–6219

Msiza IS, Nelwamondo FV, Marwala T (2008) Water demand prediction using artificial neural networks and support vector regression. J Comput 3(11):1–8

Muharemi F, Logofătu D, Andersson C, Leon F (2018) Approaches to building a detection model for water quality: a case study. In Modern approaches for intelligent information and database systems. Springer, Cham, pp 173–183

Nieto PG, Fernández JA, Suárez VG, Muñiz CD, García-Gonzalo E, Bayón RM (2015) A hybrid PSO optimized SVM-based method for predicting of the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir : a case study in Northern Spain. Appl Math Comput 260:170–187

Ocampo-Duque W, Ferré-Huguet N, Domingo JL, Marta S (2006) Assessing water quality in rivers with fuzzy inference systems: a case study. Environ Int 32(6):733–742

Oukil A, Soltani AA, Boutaghane H, Abdalla O, Bermad A, Hasbaia M, Boulassel MR (2021) A surrogate water quality index to assess groundwater using a unified DEA-OWA framework. Environ Sci Pollut Res 28(40):56658–56685

Phadatare SS, Gawande S (2016) Review paper on development of water quality index. International Journal of Engineering Research and Technology (IJERT) 5(5):765–767

Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45

Rachedi LH, Amarchi H (2015) Assessment of the water quality of the Seybouse River (north-east Algeria) using the CCME WQI model. Water Supply 15(4):793–801

Ruta D, Gabrys B (2005) Classifier selection for majority voting. Information Fusion 6(1):63–81

Saint-Jean C, Frélicot C (2001) An hybrid parametric model for semi-supervised robust clustering. In: Int. Conf. on Recent Developments in Mixture Modelling (MIXTURES)

Schölkopf B, Smola AJ, Bach F (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press

Semmlow JL (2004) Biosignal and medical image processing (Signal Processing and Communications, 22). CRC Press

Singh KP, Basant N, Gupta S (2011) Support vector machines in water quality management. Anal Chim Acta 703(2):152–162

Soltani AA, Oukil A, Boutaghane H, Bermad A, Boulassel MR (2021) A new methodology for assessing water quality, based on data envelopment analysis : application to Algerian dams. Ecol Ind 121:106952

Soltani AA, Bermad A, Boutaghane H, Oukil A, Abdalla O, Hasbaia M, Oulebsir R, Zeroual S, Lefkir A (2020) An integrated approach for assessing surface water quality: Case of Beni Haroun dam (Northeast Algeria). Environ Monit Assess 192(10):1–17

Übeyli ED (2009) Analysis of electrocardiographic changes in partial epileptic patients by combining eigenvector methods and support vector machines. Expert Syst 26(3):249–259

Vapnik V (2000) The nature of statistical learning theory. Springer-Verlag, New York

Wang LJ, Zou ZH (2008) Application of improved attributes recognition method in water quality assessment. Chinese J Environ Eng 2(4):553–556

Wang ZY, Yang YF (2010) Multi-class cluster support vector machines. J Comput Appl 30(1):143–145

Wang Y, Wang P, Bai Y, Tian Z, Li J, Shao X, Mustavich LF, Li BL (2013) Assessment of surface water quality via multivariate statistical techniques : a case study of the Songhua River Harbin region. China J Hydro-Environ Res 7(1):30–40

Wang Q, Li S, Li R (2019) Evaluating water resource sustainability in Beijing, China : combining PSR model and matter-element extension method. J Clean Prod 206:171–179

Widodo A, Yang BS (2007) Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. Expert Syst Appl 33(1):241–250

Wu CH, Tzeng GH, Goo YJ, Fang WC (2007) A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. Expert Syst Appl 32(2):397–408

Yan H, Zou Z, Wang H (2010) Adaptive neuro fuzzy inference system for classification of water quality status. J Environ Sci 22(12):1891–1896

Yang BS, Han T, Yin ZJ (2006) Fault diagnosis system of induction motors using feature extraction, feature selection and classification algorithm. JSME Int J, Ser C 49(3):734–741

Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J Hydrol 396(1–2):128–138

Zhang SW, Liu YF, Yu Y, Zhang TH, Fan XN (2014) MSLoc-DT : a new method for predicting the protein subcellular location of multispecies based on decision templates. Anal Biochem 449:164–171

Zhang W, Gao H, Sun H (2018) Application and analysis of Bayesian method and grey relational analysis in marine water quality evaluation. IOP Conf Ser Earth Environ Sci 182:012007

Zhou W, Wu B (2008) Assessment of soil erosion and sediment delivery ratio using remote sensing and GIS : a case study of upstream Chaobaihe River catchment, north China. Int J Sedim Res 23(2):167–173

Zou ZH, Yun Y, Sun JN (2006) Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. J Environ Sci 18(5):1020–1023

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Mohamed Ladjal[1] · Mohamed Bouamar[1] · Youcef Brik[1] · Mohamed Djerioui[1]

Mohamed Bouamar
mohamed.bouamar@univ-msila.dz

Youcef Brik
youcef.brik@univ-msila.dz

Mohamed Djerioui
mohamed.djerioui@univ-msila.dz

[1] LASS, Laboratory of Analysis of Signals and Systems, Department of Electronics, Faculty of Technology, University of M’sila, M’sila, Algeria