Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2022.29069

279 (2022) 61–67
December

A combined water quality classification model based on kernel principal component analysis and machine learning techniques

Smail Dilmi

Laboratory of Analysis of Signals and Systems, Department of Electrical Engineering, University of M'sila, 28000 M'sila, Algeria, email: dilmid06@gmail.com

Received 7 June 2022; Accepted 2 October 2022

ABSTRACT

Water quality monitoring plays an essential role in environmental management and the protection of water resources. However, the increasing risks of pollution make the process of monitoring using conventional methods more complex and costly. Currently, the use of automated processes based on artificial intelligence and machine learning techniques has become necessary in the field of water quality control to achieve quality control and reduce operating costs. This paper presents a comparative study of three machine learning techniques, namely K-nearest neighbors (KNN), decision tree (DT), and support vector machine (SVM), for the water quality classification of Tilesdit Dam (Algeria). Furthermore, the kernel principal component analysis (KPCA) technique was utilized to choose the important variables for water quality classification. The models were trained and tested based on historical data collected from the dam monitoring station for 3 y (2016–2018). The results of the study indicated that a combination of KPCA and DT techniques gave the best performance, with a classification accuracy of 99.68%.

Keywords: Water quality classification; Machine learning techniques; Kernel principal component analysis; Variables selection

1. Introduction

Water is one of the essential elements for sustaining life. Water quality monitoring plays a vital role in environmental management and the protection of water resources. However, the increasing risks of pollution make the process of water quality monitoring using traditional methods more complex and expensive. These methods depend on the understanding of the different descriptor parameters of clean water quality through various physico-chemical analyzes performed in the laboratory to subsequently determine its state and find the appropriate means for water treatment [1]. The disadvantage of these methods is the need for interference of a human expert for a long time to determine the state of water quality. Moreover, it is not possible to follow the determination of water quality in real-time. Thus, automation of these processes will play an important role in reducing operating cost constraints and effectively monitoring water quality in real-time.

Many approaches based on deep and machine learning techniques have been proposed for water quality assessment and classification. For example, Dilmi and Ladjal [1] proposed a new approach for water quality categorization based on a combination of deep learning and feature extraction techniques. The results of their study showed that the combination of an independent component analysis (ICA) technique with a long short-term memory (LSTM) and the linear discriminant analysis (LDA) technique with LSTM gave the best performance with a classification accuracy of 99.72%. Saghebian et al. [2] presented a model based on integrating principal component analysis (PCA)

Presented at the First International Congress of Energy and Industrial Process Engineering (ICEIPE'22) 23–25 May 2022, Algiers, Algeria

and decision tree (DT) techniques for groundwater quality classification in the Ardebil region of Iran. Their model showed a good performance for water quality classification. Sulaiman et al. [3] utilized an artificial neural network (ANN) for water quality classification. The study showed encouraging results with a classification accuracy of 80%. Shafi et al. [4] presented a comparative study between different machine learning techniques for water quality categorization. The results of their study indicated the superiority of deep neural networks over all other techniques, with a classification accuracy of 93%. Dezfooli et al. [5] investigated the performance of three machine learning techniques for classifying the water quality of the Karoon River (Iran). The results of the study showed that the probabilistic neural network (PNN) technique gave the best performance using three quality parameters with a classification accuracy of 90.70%. Prakash et al. [6] presented a comparative study of three machine learning techniques, including DT, support vector machine (SVM), and K-nearest neighbor (KNN), for groundwater quality classification in Madhya Pradesh (India). The results of the study indicated that the SVM model gave the best performance with a classification accuracy of 96.6%. Abdul Malek et al. [7] investigated the performance of seven machine learning models to classify the water quality of the Kelantan River (Malaysia). Their study showed that the ensemble model of Gradient Boosting (GB) gives the best performance with a classification accuracy of 94.90%. Nair and Vijaya [8] utilized several machine learning models to classify water quality. Their study revealed that the multilayer perceptron (MLP) model outperformed other models with a classification rate of 81%. Kaur et al. [9] presented a comparative study of three machine learning techniques for water quality classification. Moreover, the PCA technique was used to select the important variables for water quality classification. Their study showed that the combination of PCA and neural network (NN) techniques gives the best performance with a classification rate of 98.9%.

In this paper, a comparative study was presented of three machine learning techniques, SVM, DT, and KNN, for water quality classification of the Tilesdit Dam in Algeria. Furthermore, the kernel principal component analysis (KPCA) technique was employed to select the key variables for water quality classification.

2. Materials and methods

2.1. Proposed approach

Water quality monitoring and control can be considered a pattern recognition problem, where categories correspond to water quality conditions, and patterns represent the measurements of water descriptor parameters. It usually includes data acquisition, signal processing, variable selection, model learning, classification, and decision-making. Our approach is based firstly on the preparation of the database and the selection of variables important for water quality classification using the KPCA technique. Then data are entered into the classifiers (SVM, DT, and KNN).

At the level of the monitoring system, the physico-chemical parameters used can be numerous. Measurements of these parameters were converted into electrical signals from sensors installed in the water production station and transmitted to a data processing and control unity supervised by an expert for data acquisition, analysis, and decision-making. The architecture of the water quality monitoring and control system is shown in Fig. 1.

2.2. Study area and data description

Tilesdit Dam is located 122 km east of Algiers (Algeria). This dam is geographically located in the city of Bechloul,

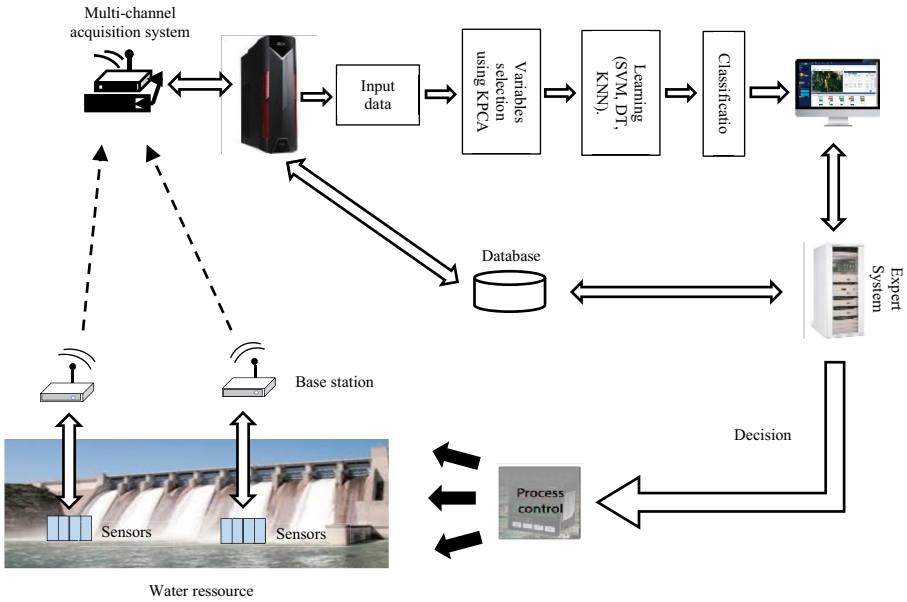

Fig. 1. Proposed system of water quality monitoring.

20 km southeast of the state of Bouira, between the following coordinates: 4° 14′ 23″E 35° 13′ 22″N. The geographical location of this dam is characterized by a semi-arid climate, that is, cold and rainy in winter, hot and dry in summer, and the average rainfall is about 440–660 mm/y.

In this work, we aim to apply our approach to classifying water quality states using measurements of physico-chemical parameters provided by some sensors installed in the plant. Measurements of these parameters were collected for a period of 3 y (2016–2018). Our knowledge of the treatment operation is exclusive to the historical data registered from this plant. Several water quality parameters are measured daily, as well as laboratory tests are performed every week at all levels of treatment. Directly after sampling, turbidity (TU), conductivity (*C*), pH, and temperature (*T*°) are measured in the area. These parameters are measured constantly (3 times a day) and at any level of the treatment operation. Then, chemical components of samples, such as full title alkaline (FTA), permanent hardness (*H*), bicarbonate (*B*), and magnesium (Mg), are analyzed every week. After the measurements are taken, they are compared to drinking standards defined by the National Water Resources Agency (NWRA) and categorized by an expert into three different classes (class one: upper, class two: medium, and class three: lower) for determining the quality of the water used. Descriptive statistics of the selected physico-chemical parameters are given in Table 1.

3. Methods

3.1. Kernel principal component analysis

Kernel principal component analysis (KPCA) is a generalization for linear principal component analysis (PCA) in the nonlinear case using the Kernel trick. The idea of KPCA is to project the data x_k through a non-linear map into the feature space $\phi(x)$ (high dimensional space) and then apply the linear PCA [10,11].

Let X be a data matrix that is composed of N observations of m variables, with $x_k \in X$. By projecting this data from the X space into the feature space $\phi(x)$ of dimension $l \gg m$, KPCA solves the following eigenvalue problem) [10,11]:

$$\lambda_i V_i = \tilde{C} V_i,\ i = 1,2..., \tag{1}$$

Table 1
Descriptive statistics of the selected physico-chemical parameters

Variables	Min.	Max.	Mean	Standard deviation
pH	7.15	8.30	7.567	0.25
C	414.00	624.00	585.393	36.278
T°	9.70	24.20	16.13	3.483
TU	1.320	23.81	3.835	2.392
Mg	7.290	47.628	22.268	4.931
B	158.620	289.14	222.497	23.213
H	0.00	168.00	32.287	23.029
FTA	130.00	237.00	181.845	18.703

where λ_i is one of the eigenvalues, V_i is one of the eigenvectors, and $C^{\sim}$ is the covariance matrix of $\phi(x)$. The evaluation of the covariance matrix in the feature space [10,11]:

$$\tilde{C} = \frac{1}{1}\sum_{k=1}^{1} \varphi\left(X_k\right)\varphi\left(X_k\right)^T \tag{2}$$

Eq. (1) can be transformed into the following eigenvalue problem [10,12]:

$$\tilde{\lambda}_i \alpha_i = K_{ij}\alpha_i,\ i = 1,2,..., \tag{3}$$

where $K_{ij} = K(x_i,x_j) = \phi(x_i)^T\phi(x_j)$ is the $l \times l$ kernel matrix. The elegance of using the K kernel is that one can compute the inner products of the transformed space without having to do the transformation explicitly. α_i represents the eigenvector corresponding to K and satisfying $V_i = \sum_{j=1}^{l} \alpha_i(j)\varphi\left(x_j\right)$. $\tilde{\lambda}_i$ represents the eigenvalue corresponding to K and satisfying $\tilde{\lambda}_i = \text{l}\lambda\text{i}$.

Finally, based on the estimated $\tilde{a}_i$, the kernel principal components (KPC) of x_k are calculated by Eq. (4) [10]:

$$S_k(i) = u_i^T\varphi\left(X_k\right) = \sum_{j=1}^{1} \tilde{\alpha}_i(j)K\left(x_j,x_k\right),\ i = 1,2,..., \tag{4}$$

3.2. Decision tree

A decision tree (DT) is a supervised learning algorithm that can be used to solve both regression and classification problems, but it is mostly preferred for solving classification problems. The DT distributes a large data set into small homogeneous data sets according to a set of discriminative variables to ensure more easy and more effective classification. Fig. 2 shows the basic structure of the DT.

3.3. K-nearest neighbors

The K-nearest neighbors (KNN) algorithm is a supervised learning method non-parametric used for classification and regression. The algorithm finds a predetermined number of training samples closest in the distance to the new sample and predicts the label from these samples based on the distance function. There are many distance

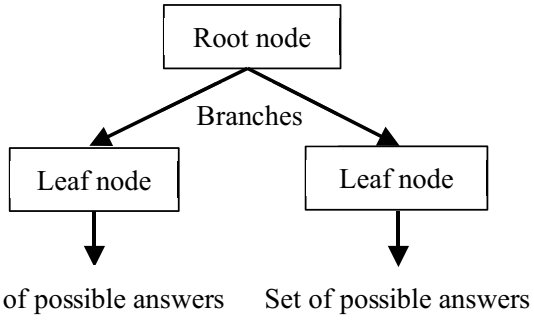

Fig. 2. Basic structure of the decision tree.

functions used in KNN, such as Euclidean, Minkowski, Manhattan, and Hamming. However, Euclidean distance is the most common option. Fig. 3 shows the working principle of the KNN approach.

3.4. Support vector machine

The support vector machine (SVM) method is one of the most popular methods in the field of machine learning for solving classification and regression problems. Its main concept is to use hyperplanes to define decision boundaries that separate data points of different categories, as shown in Fig. 4. The idea behind the SVM technique is to map the original data points from the input space into the feature space (high dimensional space) so that the classification problem becomes simpler. The advantage of the SVM technique is that it can handle both linear and non-linear classification tasks. Moreover, it can solve both binary and multi-class classification problems.

4. Results and discussion

4.1. Variables selection using KPCA

The appropriate variables are selected based on two basic conditions: first, one variable is chosen for each factor that has the highest correlation value (negative or positive); and in the second condition, the priority of selection is for parameters that can be measured directly and continuously, that is, have physical sensors such as TU, *T*°, *C*, and pH, this allows the creation of an intelligent monitoring system working continuously and permanently.

To implement this stage, a database containing 122 samples for eight physico-chemical parameters of water quality data was used. Fig. 5 shows measurements of the selected physico-chemical parameters (Tilesdit station).

From feature extraction using the KPCA technique, we can understand that there is a change in the data features to become uncorrelated components. The first three axes of the original data and the first three KPC are plotted in Figs. 6 and 7, respectively. From Fig. 7, it can be seen that the data was well separated and grouped using the KPCA technique. Table 2 represents the KPCA application for the total dataset. We can note that the first three kernel principal components represent 97.9240% of the total variation proportion:

- Axis one (57.7216%), in this axis, pH shows negative loading, whereas *B*, FTA, and *C* show strong positive loading.
- Axis two (27.6609%), in this axis, *H* and *T*° show strong negative loading, whereas pH shows positive loading.
- Axis three (12.5415%), in this axis, TU shows strong positive loading.

We can also note that the first axis correlates positively with *C* (0.7112), *B* (0.8920), and FTA (0.8822) and negatively with pH (–0.5481). However, axis two correlates negatively with *T*° (–0.7854) and *H* (–0.6988) and positively with pH (0.5135). Whereas axis three correlates positively with TU (0.7328) only. Finally, we can only keep the variables *C*, pH, and TU to build an intelligent model.

4.2. Classification results

After selecting the appropriate variables, which were *C*, pH, and TU, they were employed as inputs to the SVM, DT, and KNN machine learning systems to train the models. To implement this stage, a database containing 1,200 samples was used. In this work, the dataset was divided into 70% for training the models and 30% for testing their performance. To evaluate the performance of the models, three metrics were used; accuracy (ACC), balanced accuracy (BCA), and error. They are defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{BCA} = \frac{\sum_{i=1}^{n} \mathrm{Accuracy}}{n} \tag{6}$$

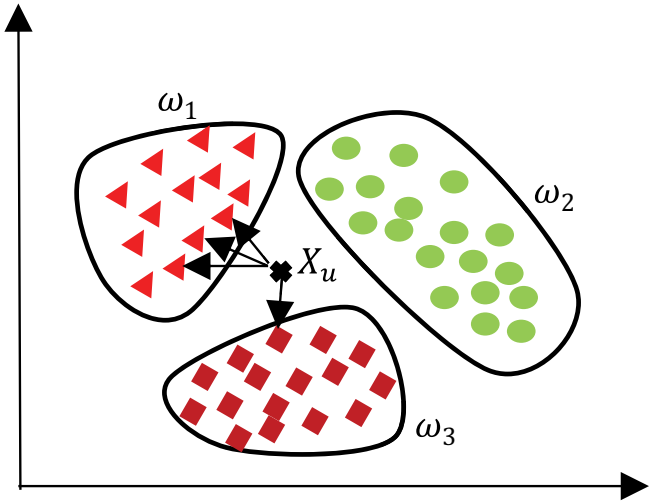

Fig. 3. KNN approach.

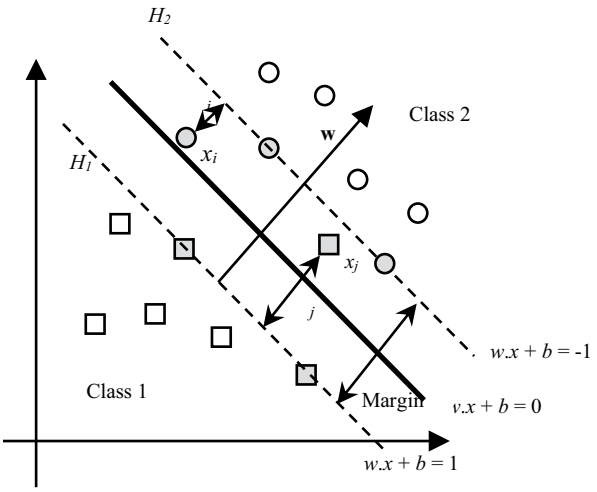

Fig. 4. Basic SVM structure.

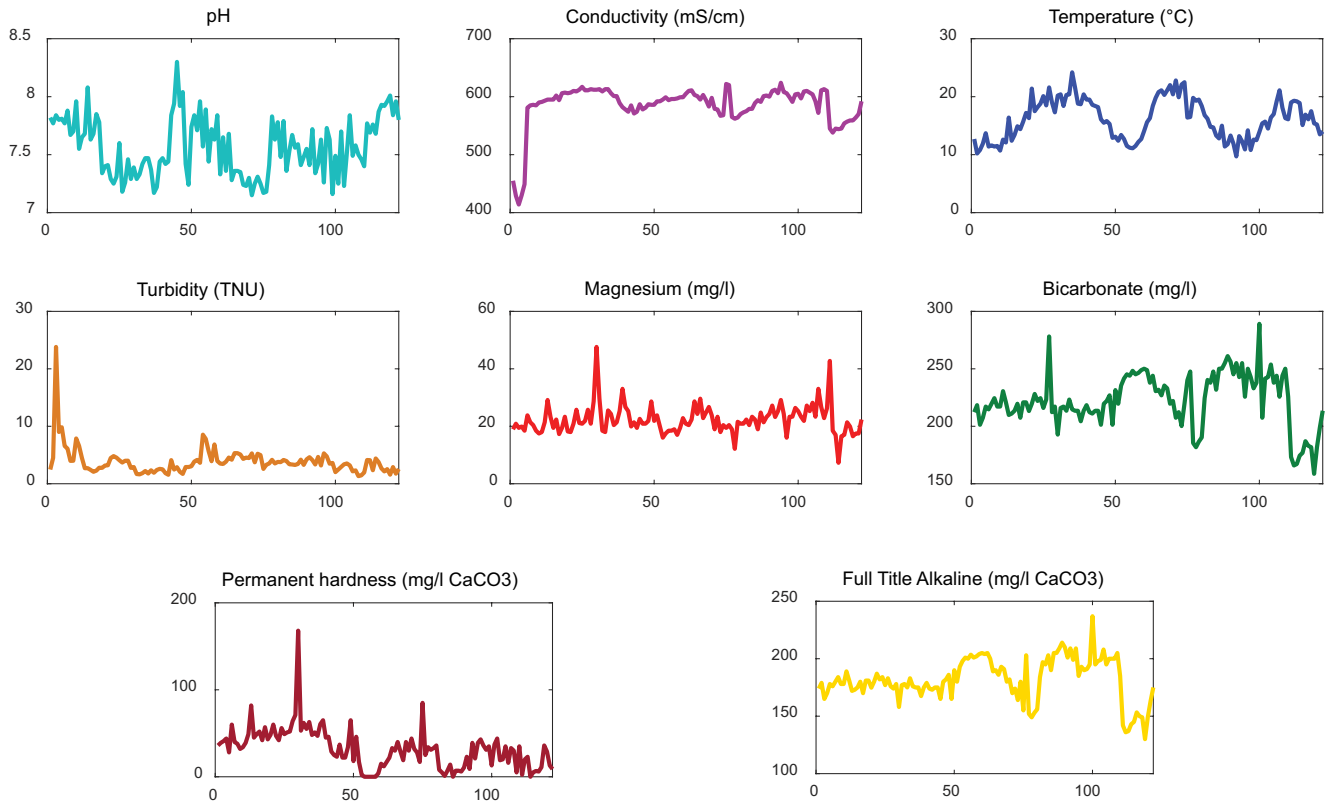

Fig. 5. Measurements of the selected physico-chemical parameters (Tilesdit station).

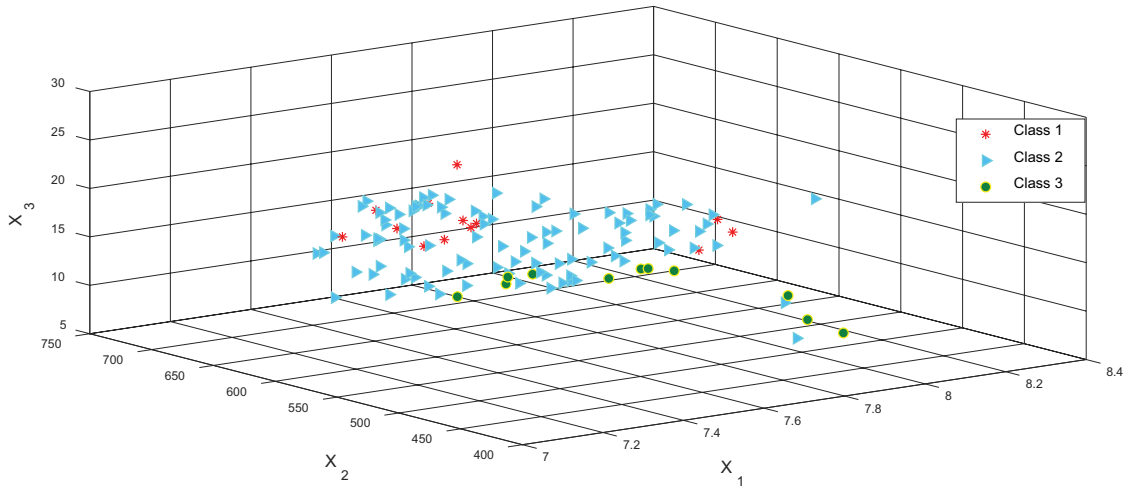

Fig. 6. Original data.

$$\text{Error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

where n represents the number of classes, and TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

Tables 3 and 4 show the results of the comparison of classification models based on the above-mentioned criteria. Table 3 shows the classification results without the selection of variables. As can be seen from the table, the DT model achieved the best performance compared to other models, with 96.88%, 88.89%, and 3.12% for accuracy, balanced accuracy, and Error, respectively. By analyzing the results of balanced accuracy, it was clear that all models suffered from instability with test data, except for the DT model, whose stability was considered to be acceptable to a large extent. Table 4 shows the classification results

Table 2
Descriptive statistics of the created KPCs

	F1	F2	F3	F4	F5	F6	F7	F8
Eigenvalues × 10^3								
	1.6031	0.7682	0.3483	0.0279	0.0174	0.0086	0.0037	0.0340
Percentage of the proportion of total variance								
	57.7216	27.6609	12.5415	1.0044	0.6278	0.3094	0.1332	0.0012
Cumulative percentage of the proportion of total variance								
	57.7216	85.3825	97.9240	98.9284	99.5562	99.8656	99.9988	100
Eigenvectors of the variables obtained using the KPCA application								
pH	–0.5481	0.5135	–0.2380	0.5020	–0.0867	–0.2893	0.1896	0.0012
C	0.7112	–0.3412	–0.3927	0.0991	0.2588	–0.3169	–0.2149	0.0007
T°	–0.0852	–0.7854	–0.1081	–0.4294	–0.2833	–0.2284	0.2174	0.0038
TU	–0.2161	0.4729	0.7328	–0.2879	0.0784	–0.3011	–0.1114	0.0080
Mg	0.3527	–0.4956	0.4239	0.4880	–0.4413	–0.0201	–0.1292	–0.0117
B	0.8920	0.3535	0.0998	–0.0314	0.0140	–0.0102	0.1994	–0.1676
H	–0.0730	–0.6988	0.4104	0.2878	0.4600	0.0144	0.2064	0.0194
FTA	0.8822	0.3939	0.0783	–0.0064	–0.0707	0.0062	0.1581	0.1737

Table 3
Performance measurements (%) for models without variable selection

	ACC (%)	BCA (%)	Error (%)
Gaussian SVM	84.38	48.22	15.62
$KNN_{K=1}$	81.25	70.67	18.75
$KNN_{K=3}$	81.25	41.98	18.75
$KNN_{K=5}$	84.38	46.55	15.62
Decision tree	96.88	88.89	3.12

Table 3
Performance measurements (%) for models with variable selection

	ACC (%)	BCA (%)	Error (%)
Gaussian SVM	98.40	72.68	1.60
$KNN_{K=1}$	96.81	84.58	3.19
$KNN_{K=3}$	98.72	92.42	1.28
$KNN_{K=5}$	98.40	85.23	1.60
Decision tree	99.68	99.87	0. 32

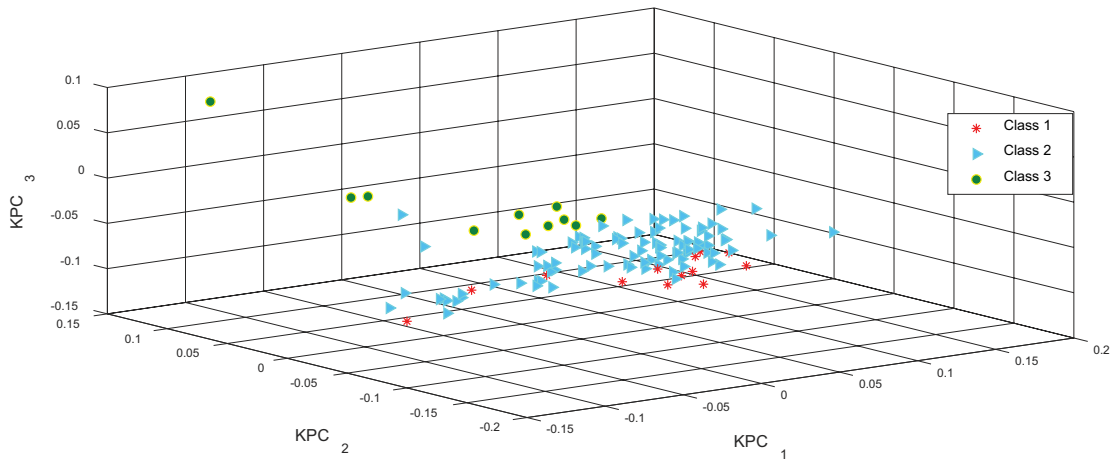

Fig. 7. Feature extraction using KPCA.

with the selected variables. As can be seen, there was a significant improvement in performance. This confirmed the importance of the step of variable selection using the KPCA technique. In general, all models achieved very satisfactory performance except for the Gaussian SVM model, which suffered from instability with the test data. On the other hand, the performance of the DT model was much better compared to the other models, with 99.68%, 99.87%, and 0.32% for accuracy, balanced accuracy, and error, respectively.

In conclusion, by analyzing and comparing the results of the study, it can be concluded that the combination of KPCA and DT techniques gave an effective approach to water quality classification.

Based on recent previous studies, the proposed approach showed superior performance over those of previous studies, such as the results of the study shown in Abdul Malek et al. [7], where their model achieved a classification accuracy of 94.90% using their proposed method. Nair and Vijaya [8], the authors utilized several machine learning models, and the MLP model gave the best performance with a classification rate of 81%. In addition, the study described in Kaur et al. [9] reported good results with a classification accuracy of 98.9% using a combination of PCA and NN techniques. In this work, a comparative study was presented of three machine learning techniques (i.e., SVM, DT, and KNN). Furthermore, the KPCA technique was employed to select the important variables for water quality classification. The combined model based on KPCA and DT techniques had the best performance, with a classification accuracy of 99.68%.

5. Conclusions

This paper presented a comparative study of various machine learning models such as KNN, DT, and SVM for the water quality classification of the Tilesdit Dam in Algeria. Moreover, the KPCA technique was used to select the important variables for water quality classification. The models were trained and evaluated based on historical data collected from the dam monitoring station for 3 y (2016–2018). The KPCA technique has allowed for keeping only three water quality parameters that were easy to measure and inexpensive, *C*, pH, and TU. The results of the study showed that the KPCA-DT model had high efficiency in terms of classification accuracy of 99.68% as well as model stability of 99.87%.

In future research work, we will (1) use soft sensors in the existence of the chemical parameters that cannot be measured directly, (2) use other nonlinear feature extraction techniques such as kernel independent component analysis (KICA) and kernel discriminant analysis (KDA), (3) test the performance of classification using new techniques of the machine and deep learning such as LSTM and convolutional LSTM (ConvLSTM).

References

[1] S. Dilmi, M. Ladjal, A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques, Chemom. Intell. Lab. Syst., 214 (2021) 104329, doi: 10.1016/j.chemolab.2021.104329.

[2] S.M. Saghebian, M.T. Sattari, R. Mirabbasi, M. Pal, Ground water quality classification by decision tree method in Ardebil region, Iran, Arabian J. Geosci., 7 (2014) 4767–4777.

[3] K. Sulaiman, L.H. Ismail, M.A.M. Razi, M.S. Adnan, R. Ghazali, Water quality classification using an artificial neural network (ANN), IOP Conf. Ser.: Mater. Sci. Eng., 601 (2019) 012005, doi: 10.1088/1757-899X/601/1/012005.

[4] U. Shafi, R. Mumtaz, H. Anwar, A.M. Qamar, H. Khurshid, Surface Water Pollution Detection Using Internet of Things, 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 2018.

[5] D. Dezfooli, S.-M. Hosseini-Moghari, K. Ebrahimi, S. Araghinejad, Classification of water quality status based on minimum quality parameters: application of machine learning techniques, Model. Earth Syst. Environ., 4 (2018) 311–324.

[6] R. Prakash, V.P. Tharun, S.R. Devi, A Comparative Study of Various Classification Techniques to Determine Water Quality, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018), Coimbatore, India, 2018.

[7] N.H. Abdul Malek, W.F. Wan Yaacob, S.A. Md Nasir, N. Shaadan, Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques, Water, 14 (2022) 1067, doi: 10.3390/w14071067.

[8] J.P. Nair, M.S. Vijaya, River water quality prediction and index classification using machine learning, J. Phys.: Conf. Ser., 2325 (2022) 012011, doi: 10.1088/1742-6596/2325/1/012011.

[9] A. Kaur, M. Khurana, P. Kaur, M. Kaur, Classification and Analysis of Water Quality Using Machine Learning Algorithms, S.K. Sabut, A.K. Ray, B. Pati, U.R. Acharya, Eds., Proceedings of International Conference on Communication, Circuits, and Systems. Lecture Notes in Electrical Engineering, Springer, Singapore 2021, pp. 389–398.

[10] L.J. Cao, K.S. Chua, W.K. Chong, H.P. Lee, Q.M. Gu, A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, Neurocomputing, 55 (2003) 321–336.

[11] A. Widodo, B.-S. Yang, Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors, Expert Syst. Appl., 33 (2007) 241–250.

[12] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a Kernel eigenvalue problem, Neural Comput., 10 (1998) 1299–1319.