Received: December 8, 2021. Revised: February 18, 2022.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Self-Organizing Maps-Based Features Selection with Deep LSTM and SVM Classification Approaches for Advanced Water Quality Monitoring

Mohamed Imed Khelil[1] Mohamed Ladjal[2]* Youcef Brik[2] Mohammed Assam Ouali[2]

[1]*LGE, Laboratoire de Génie Electrique, Department of Electrical Engineering*
[2]*LASS, Laboratoire d'Analyse des Signaux et Systèmes, Department of Electronics, Faculty of Technology, University of M'sila, PB N° 166 Avenue Ichbilia, 28000 M'sila, Algeria*
* Corresponding author's Email: mohamed.ladjal@univ-msila.dz

Abstract: Water quality control and monitoring is an important concern of countries over the world. We present in this work, the use the self-organizing feature maps of Kohonen (SOFM) as features selection technique and advanced classification techniques, such as: Long Short-Term Memory (LSTM) and Support Vector Machines (SVM). This study involved the advanced assessment of surface water quality from Tilesdit dam in Algeria. Typically, water quality status is determined by comparing collected data with water quality standards. LSTM and SVM have been applied with SOFM-based features selection for water quality classification. In this work, the training step is realized using the mentioned approaches to supervise the water quality from several physicochemical parameters. Eleven of them were collected in 4 seasons during the period (2016-2018) from study area. Experiments step using a mentioned dataset in terms of accuracy (training and test), running time and robustness, is carried out. The performance of our approach is optimized by regulating the parameter values using a SFOM based features selection method. The proposed approach outperforms current conventional methods, as this approach is a combination of strong feature selection and classification techniques. Optimal input features are selected directly from the original datasets, aiming to reduce the computational time and complexity. The impact of this result is significant both technically (lower learning time) and economically (reduced the number of sensors) and can improve obviously the performance of our monitoring system. The accuracy is more than 98% in training and testing steps with features selection process for the LSTM and SVM models. The best results of sensitivity, specificity, precision, and F-score of the two proposed models were ranged all between 96,99 % and 100%. In a nutshell, the two comparative machine learning methods provide very high classification accuracy and make a considerable solution for water quality control and monitoring.

Keywords: Water quality monitoring, Self-organizing feature maps, Features selection, Deep learning, Long short-term memory, Support vector machines, Classification.

1. Introduction

Surface water quality assessment presents a crucial and fundamental role in health protection, ecological systems, social development, agriculture, and industry, and improving environmental performance, so we should keep and protect the use of water [1]. However, conventional techniques are used to assess the water quality including: Matter Element Model, Fuzzy Synthetic Evaluation, Gray Analysis Method, Logistic Curve Model, Attribute Recognition Model and Fuzzy Logic and k-Nearest Neighbors method (k-NN) [2]. These techniques require expertise in data analysis and knowledge of water quality parameters. The above limitations can be overcome using machine learning methods so that water quality monitoring based on available sensor-generated data becomes feasible and cost-effective and his techniques are becoming very useful and popular for water quality problems [3]. The conventional techniques are unsuitable and

therefore they cannot give better performance for real-time applications as the computational time and complexity are high status with proper inaptitude due to nonlinear relationships between all modeling variables. The number of studies applying Artificial Neural Networks (ANN) and Support Vector Machines (SVM) based models that have been extensively employed in water quality monitoring has considerably increased since these recent years [3, 4]. SVMs, which are relatively new data-based learning algorithms and were introduced by Vapnik (1995), have emerged as an alternative method in ANN-dominated hydrologic research fields. Most SVM applications have been focused on surface water problems. Yoon et al. [5] 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. The statistical learning theory and structural risk minimization are the theoretical foundations for the learning algorithms of SVMs. The SVM method is considered as one of the strong and universal classifiers and approximators with a highly desired degree of accuracy in machine learning [6].

In recent years, the imperfection encourages the evolution of artificial neural networks and overcome these deficiencies. One of the limitations of classic artificial neural networks also is that, there is no memory associated with the model. Which is a problem for sequential data, like text or time series? The Recurrent Neural Network (RNN), are a commonly employed and familiar algorithm in the discipline of deep learning (DL), which was first suggested in 2006, can strongly control the deficiencies of classic artificial neural networks and exploit the deep information of data [7]. RNN addresses that issue by including a feedback look which serves as a kind of memory. So the past inputs to the model leave a footprint. It provides better results compared to traditional algorithms when dealing with real-time problems. Generally, DL offers great abilities, effectively and flexibly of learning step and contains multiple and varying nonlinear hidden layers for mapping [8]. Long Short-Term Memory (LSTM) is one of the famous architecture of DL has gained large popularity according to their high generalized performance in various fields and applications, such as: water treatment systems and hydraulic modeling [9]. LSTM networks are an extension of RNN mainly introduced to handle situations where this architecture fails. It that can memorize the previous information and applies it to the calculation of the current output. It solves the problem of gradient disappearance in traditional recurrent neural networks by selectively memorizing or forgetting some data [8], which has long-term memory capability and is suitable for processing water quality data [9]. Hence, LSTM is great tool for anything that has a sequence. It has been so designed that the gradient problem of is almost completely removed in traditional recurrent neural networks, while the training model is left unaltered. Long time lags in certain problems are bridged using LSTMs where they also handle noise, distributed representations, and continuous values. With LSTMs, there is no need to keep a finite number of states from beforehand as required in the hidden Markov model (HMM). LSTMs provide us with a large range of parameters such as learning rates, and input and output biases. Hence, no need for fine adjustments. The complexity to update each weight is reduced with LSTMs, which is an advantage.

Two major operations involved in the machine learning techniques are feature extraction and classification. The classification process is a necessary step for the decision of water quality status. For the purpose to ensure a good decision and good performance of the classifier, the preparation of data inputs is necessary and requires special treatment. The use of features selection for data inputs without any transformation to avoid redundancy has obtained significant attention [10]. The advantages of features selection include a reduction in the amount of data and execution time to achieve the learning phase, avoiding redundancy, reducing of sensors in the monitoring system, low cost, improving classification accuracy.

The water quality data analysis seems a difficult task because the data have complex interactions with each other, multidimensional, changeable, and non-linear [9]. Statistics-based diagnosis techniques are preferable for implementing knowledge extraction in water quality data. Multivariate analysis methods, such as self-organizing features maps of Kohonen (SOFM) is an unsupervised learning method is used as features selection technique [11]. It can also be used to analyze and deal with non-linear, noisy, redundant, irregular, or missing and outlier data sets, excellent visualization capabilities purposes and interpretation in the category, internal relationships of samples and their index, clustering and noise reduction and classification. SOFM belongs to a kind of statistics-based diagnosis technique and has been broadly

used in the initial steps of water quality assessment applications and studies [11, 12].

LSTM has a good potential to achieve effective data representation for building improved classification tasks with SVM. SOFM helps rebuild input representation and converts it to reduced feature representation of data related to the input data, thereby improving the performance of the classification task considerably. The main contributions of this work are as follows:

(1) We used a novel deep learning approach based on the LSTM framework with SVM. We study the potential of our proposed approach to achieve effective representation and dimensionality reduction using SOFM method for the improvement of the binary classification results of shallow and traditional supervised machine learning algorithms.

(2) Better or at least similar and competitive results are achieved compared with the results of similar approaches. Moreover, our approach considerably reduces training and testing times.

In this paper, a comprehensive approach using LSTM and SVM classification methods for the decision of water quality status in combination with SOFM based features selection is proposed. The problem is regarded as a classification based on two classes of water quality (drinkable state and not drinkable state) performed on physicochemical parameters. The classification techniques are employed in a comparative study to show the performance evaluation of monitoring models, and to get the best decision and adequate solution in terms of accuracy (training and test) and running time.

The rest of paper is organized as follows. The next section describes study area and brief description of the dataset collection. Section 3 briefly describes the proposed framework for water quality monitoring and methods used and Section 4 presents and discusses the results. Finally, the last section is dedicated to conclusions and future research.

2. Study area and datasets

The surface water of Tilesdit dam is geographically located about 20 km southeast of the city of Bouira and 122 km east of Algiers (Fig. 1). It is situated between the geographic coordinates: 36° 13’ 22’’N 4° 14’ 23’’E [13].

In this research, physical sensors installed in the Tilesdit production station provided us with several important physical and chemical parameters. Our mission was limited during three years (2016-2018) to know the different treatment process and to collect data from this station. Descriptive measures of the used water quality parameters are shown in Table 1.

Table 1. Descriptive measures of the used water quality parameters

Parameters	Min	Max	Average	Standard deviation
pH	7,32	8,36	7,87	0,18
C	515,00	605,00	568,59	33,47
T°	9,80	16,00	12,16	1,28
TUR	2,81	10,40	4,93	1,81
Mg2+	7,29	47,63	22,69	5,16
TH	122,00	326,00	215,14	24,09
NH4+	0,00	0,49	0,13	0,11
NO2-	0,00	0,46	0,07	0,08
TDS	287,00	303,00	294,19	4,48
DO	3,28	5,75	4,31	0,53
PEH	0,00	168,00	31,49	22,37

Figure. 1 The Tilesdit dam – Bouira – Algeria [Google Maps]

3. Proposed framework for water quality monitoring

The goal of this study is to establish decision support models for an advanced water quality monitoring through the installed sensors for data acquisition. Generally, this advanced system includes the particular hardware and computer software, such as: sensors, conditioning circuits, data acquisition, wireless communication, signal processing blocks for large datasets, and be complemented by input data preprocessing using pre-trained SOFM model based features selection, and final decision using LSTM and SVM classification methods. All these steps are performed to obtain the most successful solution for water quality assessment. In the next sections, a brief demonstration of the principal techniques used in this work, is provided.

3.1 Features selection based on self-organizing features maps

Self-Organizing Features Maps of Kohonen (SOFM) is a kind of nonlinear neural-network model for multivariate data analysis. The SOFM is topology-preserving mapping from input data space $\Re^n$ onto a regular or hexagonal two-dimensional array of nodes like a dimensionality reduction technique (Fig. 2(a) and 2(b)) [11].

A weight vector correspond to the input vector $m_i \in \Re^n$ (prototype or reference vector) is associated to every node *i*. It is selectively optimized for better learning performances. Initially, weights are randomly distributed, and over much iteration, the SOFM eventually settles into a map of stable sectors. Each input data vector $x \in \Re^n$ is compared to the prototype vector m_i, and the best match m_c defines the winning reference vector [11].

Each sector is effectively a feature classifier. The input x is then mapped onto the corresponding location on the hexagonal two-dimensional grid mi in our case (gray symbols in Fig. 2(b) and 2(c)).

The SOFM is an unsupervised learning method. At each learning time *t,* a data pattern $x(t) \in \Re^n$ is mapped to the corresponding location in the grid, where the node *c* is that best represents the input data sample for using (matching unit BMU), and the reference vectors m_i in the neighborhood of BMU unit are moved to the selected vector $x(t)$. $c(t)$ as shown by the following Euclidean distance [14]:

$$c(t) = arg \ \min_i \{\|x_t(t) - m_i(t)\|\} \quad (1)$$

The node *c*, as well as the best neighboring units, is trained to more precisely represent the incoming data sample. The following training rule (Eq. 2) is used to update the weight vector of unit *i* [15]:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t)(x(t) - m_i(t)) \quad (2)$$

where h_{ci} is a gaussian neighbourhood function towards BMU unit, signifying how much unit *i* is updated when unit *c* is the winner. $\alpha(t)$ is the training rate decreasing with time. The neighbourhood function (Eq. (1)) typically is a symmetric around the location of the winner, monotonically decreasing function of the distance between nodes *i* and *c* on the map network [11]:

$$h_{ci}(t) = exp(\frac{\|r_c - r_i\|^2}{2\delta(t)^2}) \quad (3)$$

where $\| \ \|^2$ denotes the distance between units *c* and *i* in the output space, r_c signifies the two-dimensional location vector of unit *i* in the grid, and δ represents the time-varying parameter that controls the reduction of the neighborhood kernel function during the learning phase [12].

From *i* iterations of Eq. (2), the weight vectors of neighbouring units corresponding to different inputs become gradually similar due to the neighbourhood function h_{ci}, eventually leading to global ordering of the model vectors [11]. The amount of model vector movement is guided by a training rate α. With time, the m_i then tends to become ordered along the array in a significant way.

The SOFM can be used to visualize and interpret datasets on different group, present internal relationships of samples and their index [12]. The prototype distance between neighborhood units is showing by U-Matrix (Fig. 3(a)). It can be used as well for clustering, noise reduction and classification.

The component planes (CPs, base map of Fig. 3(b)) in SOFMs can be used to reduce redundancy in the data space (correlation hunting). A CP is built on the trained SOFM (*i* units) where each unit *i* is represented by a particular component of the corresponding reference vector m_i.

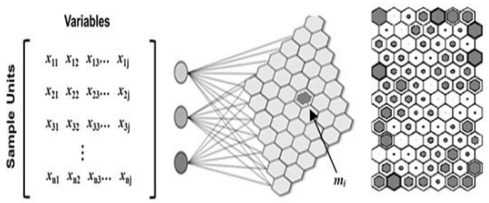
The components of the absolute correlation matrix A between all components is given as [14]:

$$a_{ij} = \frac{1}{N} \sum_{n=1}^{N} \left\| m_{ni} . m_{nj} \right\| \quad (4)$$

As input data, the correlation matrix of the component planes prototypes can be used for the learning step of a second SOFM on a rectangular grid map. The input data samples vector $x(t)$ is defined by [11]:

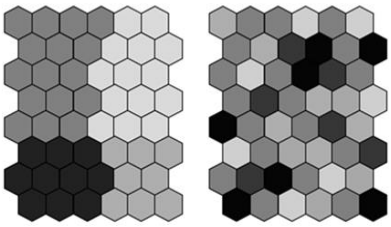
$$x_t(t) \stackrel{\text{def}}{=} a.j \quad (5)$$

where a·*j* represent the component planes (CPs).



(a) (b) (c)

Figure. 2 Principale of the SOFM: (a) Data and prototypes, (b) SOFM network, and (c) SOFM layout and hits



(a) (b)

Figure. 3 The trained SOFM (black stands for high distances): (a) U-matrix and (b) Component planes

3.2 Long short term memory deep learning

Long short term memory (LSTM) model is a new kind of the standard recurrent neural networks (RNN) by adding memory blocks called *cells* with a unique method of communication and is ordered in the form of a chain structure. It was introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [16]. LSTM network has the most essantial three gates (or layers) of model used to control the memorizing process avoiding the long-term dependency problem, including: forget, update and output gates layers that used to update of information of input gate layer (weights) contained in cell memory status. This last one (cell - memory blocks) is the fourth layer and is considered as an important element of LSTM model (Fig. 4).

The output of LSTM dependent on previous and current neuron inputs and weight with feedback at each neuron. LSTM contains internal loops that maintain useful and correct information and abandon detritus, to overcome the vanishing gradients problems caused by a long and correlated data samples. The data sample can be added or removed to the cell memory state through sigmoid function. LSTM can be described by the following formulations [9, 16, 17] :

-Forget gate layer :

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (6)$$

where σ is the logistic sigmoid function. W and b are the weight matrices and bias, respectively of the forget gate f_t. x is the input. h is hidden cell memory vectors.

-Update or input gate layer:

$$u_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (7)$$

were W_i is the weight matrices and b_i is the bias in the update gate u_t.

-New memory cell:

$$\tilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (8)$$

were W_c is the weight matrices and b_c is the bias in the new memory cell $\widetilde{C_t}$.

-Final memory cell:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (9)$$

where C_t is current cell status value, C_{t-1} is the last time frame cell status value and $\widetilde{C_t}$ represent the update for the current cell status value.

-Output gate layer:

$$o_t = \sigma(W_0.[h_{t-1}, x_t] + b_0) \quad (10)$$

$$h_t = o_t \times tanh(C_t) \quad (11)$$

were W_0 is the weight matrices and b_0 is the bias in the output gate o_t.

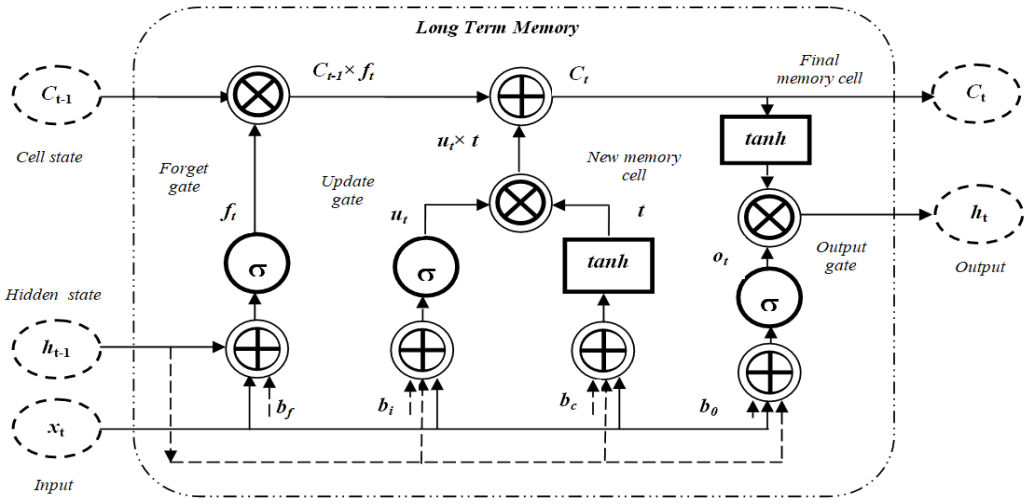


Figure. 4 The structure of the LSTM network

The forget gate f_t computed in Eq. (6), is similar to those of the update gate. It allows the model to choose when to forget the information in cell unit (cell memory) C_t. The update gate u_t, as shown in Eq. (7), help the model to decide when u_t updates C_t by using a sigmoid function σ; (σ = 1: $\widetilde{C}_t$ is updated, and σ = 0: $\widetilde{C}_t$ is ignored). Eq. (8) computes a candidate$\widetilde{C}_t$. The *tanh* function gives weight to the values which passed by, deciding their level of importance (-1 to 1). $\widetilde{C}_t$ and u_t update the parameters to new cell state C_t using Eq. (9), new cell state C_t is computed and transferred to the next layer by multiplying the update gate with the candidate $\widetilde{C}_t$ and adding it to the forget gate that is multiplied with the previous cell state C_{t-1}. The output gate o_t is computed in Eq. (10) by multiplying current input with weight W_0 and add bias b_0. Finally, the output of the sigmoid gate (o_t) is multiplied by the new values created by the tanh layer from the cell state (C_t), with a value ranging between -1 and 1 and the result h_t to send it to the next layer to keep tracking the parameters using Eq. (11).

With suitable parameters adjustment, the result value h_t is calculated according to Eqs. (9) and (10) based on C_t and o_t values. All weights of the four gates are updated based on the difference between the output and the actual values following back-propagation through time (BPTT) algorithm [17].

3.3 Support vector machines

The SVM method was introduced by Vapnik for classification, regression and density estimation [18, 19]. In SVM, all the input patterns can be separated by a linear optimal hyperplane (Fig. 5).

It is implemented through maximization of the margin around a hyperplane by mapping through some linear or non-linear functions (kernels functions) into the high dimensional features space. The mapping and maximization of the margin is determined to provide more performances and ability of generalization. The maximization of margin around a hyperplane is defined as a quadratic optimization problem.

In this case, the binary quadratic classification problem is established from the following dataset:

$$(x_i, y_i), y_i \in \{-1, +1\}, i = 1, \ldots, n \qquad (12)$$

$x \in \Re^d$, n is the number of samples and y_i is the corresponding output label (class).

The maximization of the margin around a separating hyperplane is a non- linear quadratic optimization problem using the Lagrange multipliers

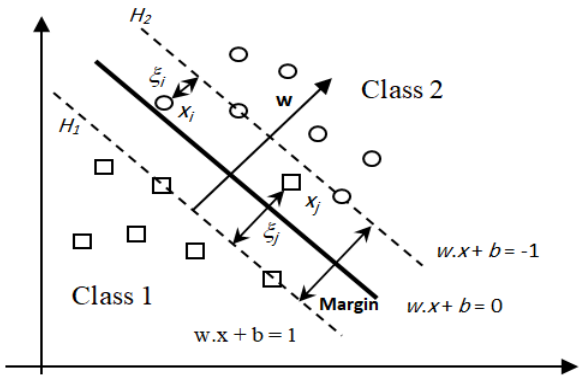


Figure. 5 The structure of SVM hyprpalne and margin

αi and non-linear mapping via a kernel function, the Lagrangian dual problem is becoming [13,20]:

$$\begin{cases} \max\limits_{\alpha_i} & L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{with} & \sum_{i=1}^{n} \alpha_i y_i = 0 \,, 0 \le \alpha_i \le C \,, i = 1, \ldots, n \end{cases} \qquad (13)$$

According to necessary and sufficient condition of *Karush-Kuhn-Tucker (KKT*), an optimal α, is given by [20, 21]:

$$\alpha_i^0 \{y_i[(w_0 x_i) + b_0] - 1\} = 0, i = 1, \ldots, n \qquad (14)$$

$$SVs = \{x_i \;\; \text{that} \;\; \alpha_i \succ 0\} \qquad (15)$$

SVs are the Support Vectors which participate in the construction of the optimal hyperplane.

The non linear decision classification function is defined by [10,20]:

$$f(x) = sign(\textstyle\sum_{SVs} \alpha_i y_i K(x_i, x) + b) \qquad (16)$$

If $f(x) < 0$, then x is belongs of class -1; if not, it is belongs of class 1, and b is the solution.

It can be used any positive definite kernel function that satisfies Mercer's theorem. The best settings of the appropriate function and its parameters are very important to efficient classification. The most types of SVM kernel functions are: [3, 20, 21]:

The polynomial function:

$$K(x, x') = (\gamma x^T x' + c)^d \quad with \;\; \gamma, c \ge 0 \;\text{and}\; d \in \mathbb{N} \qquad (17)$$

The Gaussian RBF function:

$$K(x, x') = exp(\frac{\|x - x'\|^2}{2\sigma^2}) \qquad (18)$$

4. 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 two steps: features selection and recognition of the water quality status. The feature selection technique is based on SOFM, and classification technique is based on LSTM and SVM. The hardware used to perform our simulation experiments are as follows: we have used an Intel Core TM i7-6820HQ and 2.71GHz 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.

4.1 Features selection

Feature selection for dimensionality reduction is a popularly used step in machine learning to ensure a good decision, good performance and high level of classification ability of the fitted model. The first analysis is the process of selecting a subset of the relevant and uncorrelated variables using SOFM to determine the input data for the proposed classifier models. Also, we can understand that there is a change from data features to components which are uncorrelated. In machine learning, the reduced features space is usually performed by constructing subset of the new dimensions using SOFM method.

In this step, the data base used for this work consists of 1800 samples from eleven physico-chemical parameters of water quality (Table 1), these parameters are: *Temperature* (T°-°C), *Conductivity* (C-µs/cm), *pH*, *Turbidity* (TUR-NTU), *Dissolved Oxygen* (DO- mg/l), *Magnesium* (Mg2+-mg/l), *Total Hardness* (TH-°F), *Permanent Hardness* (PEH-°F), *Total Dissolved Solids* (TDS), *Ammonium* (NH4+-mg/l) and *Nitrite* (NO_2^- mg/l).

The proposed approach of features selection is used to identify the correlation and similarity of input data parameters by employing the visualization power and clustering of SOFM. In experiments step, a [6 × 10] SOFM composed of 60 units (neurons) is used for features selection. Due to the topological preserving property of SOFM, the input data parameters that have close proximity with each other are mapped to the same neurons or its neighbors arranged on the map network. The distance between the reference or prototype vector of neuron and its neighbors is calculated by U-matrix (the unified distance matrix). Fig. 6 visualize by using colour scale, the distribution of all variables or component planes on the SOFM for the input data vectors and the reference vector distances between nodes in U-matrix plane. The larger distance is plotted in yellow color. For example, the objects with high values for Dissolved Oxygen are located in the downright part of the SOFM plane. The component plane gives some information visually about the relation between a parameter and the clusters. The parameters like: Conductivity, TDS and Dissolved Oxygen change from small values (left-bottom area) to big values (right-bottom area) on the map. pH had a graduation of the color almost too light carrying the values higher on the right. In order to group the trained SOFM units, it is advantageous to use the U-Matrix algorithm which has led to the identification of 3 clusters located on the left, central and right side. However, it is difficult to recognize the influence of indicators definitely from the component planes.

The selecting of the subset of the component planes (Fig. 6) using SOFM shows four different groups of variables. The first subset group is formed by the following parameters: Ammonium, Nitrite and Turbidity. The second subset group includes the parameters such as: Total Hardness, Permanent Hardness, Magnesium and Temperature. The third subset group is formed by the parameters like: Conductivity, Total Dissolved Solids and Dissolved Oxygen. The last group is constituted by pH. Using the ordering planes, a proper selection of surface water quality parameters could be done. Each well-defined group could be selectively presented by one of its members. Thus, Turbidity and Conductivity were selected to represent the first and the second group, respectively. Conductivity was selected to represent the third group. The last selected variable pH represents fourth group. The selected water quality parameters could be more reliably and accurately analytically determined and are directly related to specific anthropogenic influences along the dam catchment.

In the presence of the chemical parameters that cannot be measured continuously. So the variables retained are: *Conductivity* (EC), *Temperature* (T°), *Turbidity* (TU) and *pH*. The four parameters are selected as input of the proposed monitoring models that are more easily measured continuously. This solution is in any case not final, it will probably be necessary to perform a periodic re-learning system to take account of situations likely to be encountered and to allow continuous adaptation of it to any changes in the water quality. The new data set of 1800 samples described by the four selected water quality parameters was further subjected to chemometric treatment by statistical learning approaches.

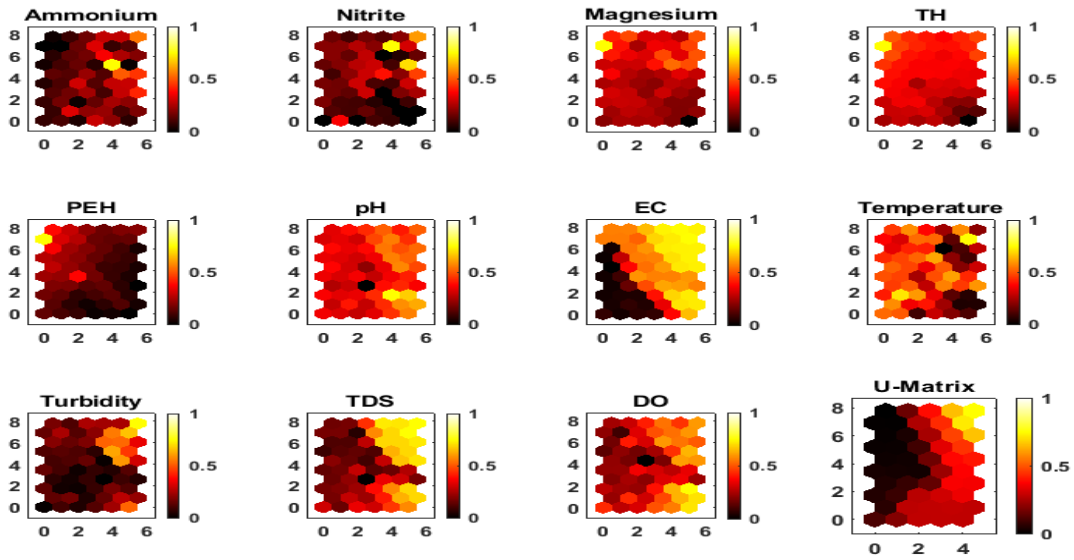


Figure. 6 Visualization of variable planes (CP) and U-matrix for the input data displayed on SOFM

The goal is to predict the water quality status from uncorrelated selected variables with lower global cost system and better quality.

4.2 Training and classification

In this section, we evaluated the proposed framework on several physicochemical parameters used for water quality decision. According to the drinking water Quality guidelines, two classes of water quality have been considered: (Class I: *upper*, Class II: *lower*). In order to proceed with the tests, training and test sets constituted of real data relating to the various qualitative water statuses are used.

In this application, two diverse base classification models, including LSTM and SVM are employed for water quality assessment.

The training data of 1800 samples were collected from Jan. 1, 2016 to Dec. 31, 2018, all constituted of the four physicochemical variables: *Temperature* (*T*°), *pH*, *Electrical Conductivity* (*EC*) and *Turbidity* (*TU*) reduced by SFOM-based features selection step as input of the proposed classification models.

4.2.1. Evaluation metrics

Through trial and error experiments, the performance metrics in training and testing phases such as accuracy (*Acc* : *the percentage of predictions those are correct*) with features selection, Sensitivity (Recall, *Se : True Positive Rate*), Specificity (*Sp*), Precision (*Pr: Positive Predictive Value*), F-score (*F-Scr*) and training time (*T_train*) were used to measure the success of the proposed model. The values of these metrics are calculated by the following Equations [22, 23]:

$$\begin{aligned} Acc\ (\%) &= \frac{Number\ of\ Correct\ Classifications}{Total\ Number\ of\ Test\ Samples} \\ Pr &= \frac{TP}{TP+FP} \times 100\% \\ Se &= \frac{TP}{Pr\frac{TP}{TP+FN}} \\ Sp &= \frac{TN}{TN+FP} \times 100\% \end{aligned} \tag{19}$$

The true positive (*TP*) and the true negative (*TN*) correspond to correctly classified samples of each class, whereas the false positive (*FP*) and false negative (*FN*) correspond to incorrectly classified samples, respectively.

4.2.2. Testing the proposed hybrid feature selection and classification methodology

To evaluate the proposed methods, standard 10-fold cross-validation and simple Holdout has been implemented in training and testing phases. To validate the generalization ability of the proposed models with simple Holdout, the raw data are divided into two subsets: 60% is used for training, and the remaining 40% is used for testing. The experiment results with all performances metrics of each individual proposed method are presented.

The final LSTM architecture used in this paper consist of an input layer followed by an LSTM layer. The network ends with three connected layers, a fully connected layer, a softmax layer and output layer to classify the water status. Each hidden layer are followed by a dropout layer implemented between hidden layers and applied after each LSTM layer for regularization (dropout value = 0.2). In this application, various architectures of LSTM have been tested to determine the adequate number of

Table 2. The results of the LSTM classification model in the training and testing phases

Models	Performances metrics					
	Pr	***Se***	***Sp***	***F-Scr***	***T_tr (s)***	***Acc***
LSTM_Holdout (Training)	99.64%	99.02%	96.99%	99.33%	1.2863e+03	98.81%
LSTM_Kfold (Training)	100%	97.44%	100%	98.70%	936.03	97.67%
LSTM_Holdout (Testing)	99.17%	99.37%	93.44%	99.27%	0.09	98.70%
LSTM_kfold (Testing)	100%	97.95%	100%	98.96%	0.08	98%

hidden layers and the neurons in hidden layers. The hidden layer consists of 100 neurons, and select 0.005 as the learning rate with batch stochastic gradient descent (batch_size) of 72 time-steps as an optimizer in order to equilibrate the convergence speed and accuracy using a standard grid search for 50 epochs was selected by trial and error. The last layer (classification output) of the LSTM is a dense layer and the loss function is the mean square error function. Furthermore, the LSTM is compared with SVM with SOFM-based features selection. Table 2 show results associated with LSTM models in training and testing phases using water quality data input. We apply standard *k*-fold cross-validation and simple Holdout procedures to evaluate the performance metrics of LSTM models for water quality classification.

The Table 2 presents the result of LSTM technique. The LSTM-SFOM hybrid model with Holdout cross validation method gets an accuracy of 98.81% (training) and 98.70% (testing), and the LSTM-SFOM hybrid model with *k*-fold cross validation method gets an accuracy of 97.22% (training) and 98% (testing).

The SVM is applied to perform the classification process using the *Gaussian RBF* and *polynomial* kernel functions. There are two parameters associated with these kernels: C and σ. In addition, polynomial kernel has also a parameter d related to the polynomial degree. Furthermore, the variation of kernel function is performed to show the excellent of characteristic of kernel function and its performance in classification process. Therefore, improper selection of parameters C, σ and d can cause over fitting or under fitting problem [13]. The goal of this guideline is to identify optimal choice of these parameters so that the classifier can accurately classify the data input using *k*-fold cross-validation. Basically, all the pairs of (C, σ) for RBF kernel and (d, C, γ) for polynomial kernel are tried and the one with the best cross-validation accuracy is selected.

Table 3 and 4 shows results of training and testing steps corresponding to SVM multi-class models. The parameters, such as: kernel function, the recognition rates for training and testing phases and different performance metrics are indicated for various values of the factor C with the linear, polynomial and radial basis kernel functions.

In Table 3 and 4, the recognition rate with SFOM features selection ranged from 98.10% to

Table 3. Water quality classification using SVM-SFOM model and selected kernel parameters in the training phase

Models	Kernel Parameters		Performances metrics					
		NSV	Pr	Se	Sp	F-Scr	T_tr (s)	Acc
SVM_Holdout	Linear ($d = 1, \gamma = 3.53, C = 879.23$)	31	97.95%	99.91%	85.80%	98.92%	771.67	98.10%
	Polynomial ($d = 2, \gamma = 37.17, C = 30.03$)	34	99.73%	99.73%	97.86%	99.73%	926.89	99.52%
	Gaussian RBF ($\sigma = 91.54, C = 913.48$)	70	100%	99.12%	100%	99.56%	71.98	99.21%
SVM_KFold	Linear ($d = 1, \gamma = 1, C = 0.007$)	29	99.90%	99.81%	99.42%	99.85%	311.76	99.75%
	Polynomial ($d = 2, \gamma = 12.07, C = 0.72$)	28	99.90%	99.71%	99.42%	99.81%	575.37	99.67%
	Gaussian RBF ($\sigma = 50.60, C = 957.51$)	41	99.90%	99.71%	99.42%	99.81%	575.37	99.67%

Table 4. Water quality classification using SVM-SFOM model and selected kernel parameters in the testing phase

Models	Kernel Parameters	Performances metrics					
		Pr	Se	Sp	F-Scr	T_ts (s)	Acc
SVM_Holdout	Linear (d = 1, γ = 3.53, C = 879.23)	97.95%	99.91%	85.80%	98.92%	771.67	98.10%
	Polynomial (d = 2, γ = 37.17, C = 30.03)	100%	99.79%	100%	99.90%	0.002	99.81%
	Gaussian RBF (σ = 91.54, C = 913.48)	99.79%	100%	98.36%	99.90%	0.003	99.81%
SVM_KFold	Linear (d = 1, γ = 1, C = 0.007)	99.65%	99.83%	92.86%	99.74%	0.002	99.50%
	Polynomial (d = 2, γ = 12.07, C = 0.72)	99.30%	99.82%	86.67%	99.56%	0.002	99.17%
	Gaussian RBF (σ = 50.60, C = 957.51)	96.86%	99.64%	58.14%	98.23%	0.002	96.67%

99.75% in training step and among 96.67 % until 99.81% in testing step for the SVM models. The recognition rate in training phase for linear kernel is usually lower than Polynomial and Gaussian RBF kernel. For SVM model, the performance of classification process is increased due to the feature's selection, because of SFOM searches the uncorrelated components from the input data space and treat it so that more useful in classification.

Therefore, the effect of selection of kernels functions and its parameters C and σ or d and γ is very important to achieve a good performance in training and testing sets and there are no definite rules governing its choice that might yield a satisfactory performance. Indeed, the performance depends on the choice of these parameters by the use of optimization process via cross validation methods.

4.2.3. Comparison of classification performance of classifiers

In this study, the main contribution is to train and classify the water quality data using a LSTM compared to SVM with SFOM based features selection. Compared with ordinary neural networks, LSTM has repetitive neural network modules, which can automatically process the key semantic information of the input data. In this study, the optimization of activation function and post-processing optimization is carried out based on the original LSTM network, which greatly improves the accuracy of classification results. The success of the proposed framework performance can be seen in Table 2, 3 and 4. To the best of our knowledge, it is among our contribution to propose the use of SFOM with LSTM techniques in field of water quality monitoring. For more classification performances, we used two methods of cross-validation namely: standard k-fold and Holdout, to estimate the performance of the used classification models. It appears that on the decisional level, the two models perform good results with recognition rates more than 98% in training and testing steps with features selection process. As a best result of the training on LSTM-SFOM model, 98.81% accuracy, 99.02% Sensitivity, 96.99% Specitivity, 99.64% Precision, and 99.33% F1-Score performances metric values are obtained. The SVM-SFOM model with the proper pair (d = 2, γ = 37.17, C = 30.03), has 99.81% accuracy, 99.79% Sensitivity, 100% Specitivity, 100% Precision and 99.90% F1-Score metric performances. The SVM-SFOM model requires less time than the training and testing time of LSTM-SFOM model. The longest training period has belonged to this last model. The classification results clearly showed higher accuracy and sensitivity for the two machine learning methods. According to the accuracy in the Table 5, the proposed method showed least similar and competitive results with an accuracy rate of 99.81% compared to Mesut Togaçar et *al*. [24], Ahmed et *al*. [28] and Djerioui et *al*. [29] using the Auto-Encoder deep learning method, ELM and SVM with Water quality database.

This finding is in agreement with the results published by several works. They have reported that the LSTM or SVM models have high classification accuracy for water quality data, as it makes use of the advantages of these methods that address the shortcomings of conventional techniques with different methods, databases, and fields when considering the obtained overall performance. In

Table 5. Comparison between the proposed methods with earlier reported classification methods in different dataset

Study	Data	Methods	Accuracy
Sourav Kundu et al. [25]	EEG signal	SAE-ESVM	95.5%
Indu Saini et al. [26]	MIT-BIH Arrhythmia database.	KNN algorithm	99.81
Shao Haidong et al. [27]	Fault diagnosis	Enhancement Auto-Encoder	87.8%
Ahmed et al. [28]	**Water quality database**	**MLP**	**85%**
Djerioui et al. [29]	**Water quality database**	**SVM and ELM**	**99,3%**
Mesut Togaçar et al. [24]	**Waste water quality database**	**Auto-Encoder**	**99.95**
This study	**Water quality database**	**SVM**	**99,81%**
		LSTM	**98,70**

this work, we find better or at least similar and competitive results are achieved compared with the results of similar approaches. Moreover, our approach considerably reduces training and testing times by using SOFM based features selection. The impact of this result is significant both technically (lower learning time) and economically (reduced the number of sensors) and can improve obviously the performance of our monitoring system.

The results obtained emphasize the explanation of the theoretical and practical reasons in the introduction section for the use of LSTM technique for this type of application compared to other existing techniques. Unlike some other existing techniques, and especially CNN, Auto-Encoder deep learning, ELM, KNN, MLP, SVM and other combined models that is designed to use spatial information in data, LSTM is developed to work differently. Usually, LSTM is used to process and make predictions given sequences of data. The main advantage of LSTM is its ability to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Furthermore, LSTM models can predict future values based on previous, sequential and big data for providing greater accuracy with better decision-making. In our case, the data we used for water quality assessment is presented with different seasonality which makes LSTM deal well with it compared to other conventional machine learning models.

5. Conclusions

In this work, we introduces a robust classification framework of water quality status using LSTM and SVM machine learning combined with SFOM-based features selection. The study area is the Tilesdit dam from Algeria. In this study, the proposed framework is examined in two sections as feature selection and classification. The proposed methods have obtained maximum accuracy and achieve higher and acceptable performance than existing classifier approaches. The addition of the SOFM algorithm in the proposed framework has further improved the classification accuracy. This result is important because it reflected an economic impact on the overall cost of the monitoring system (off line and reduced number of sensors). The water quality parameters obtained from the features selection process have been successfully classified by the two machines learning methods with accuracy more than 98% and metric performances more than 96.99% and reached at 100% in training and testing steps. The proposed framework has outperformed conventional techniques in terms of accuracy, computational speed and other performance metrics. The proposed approach especially based LSTM with SFOM model can be applied for any real-time monitoring and big data application with the help of appropriate sensors that it reflected an economic impact on the overall cost of the monitoring system of water quality. In future work, a novel hybrid proposed CNN-SVM or LSTM-SVM models architectures and all high-performance deep learning models will be realized in real-time recognition in practice. The single CNN architecture, feature extraction and feature classification techniques are combined into a single model. The accuracy of the system decision can be improved by exploiting new input parameters, or by using soft sensors in the presence of the chemical parameters that cannot be measured continuously.

Acknowledgments

The authors would like to thank all engineers of Tilesdit dam direction for providing the free access to databases. They would also like to thank the editor and reviewers for helpful suggestions and constructive comments. This work is supported by the DGRSDT, Ministry of Higher Education and Scientific Research of Algeria.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Author Contributions

All authors contributed to the study's conception and design. All authors read and approved the final manuscript.

Mohamed Imed KHELIL: Software, Validation, formal analysis, Investigation, Resources.

Mohamed LADJAL: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing-original draft preparation, Writing-review and editing, Visualization, Supervision, Project administration.

Youcef BRIK: Software, Validation, Formal analysis, Investigation, Visualization.

Mohammed Assam OUALI: Writing-review and editing, Validation, Formal analysis, Investigation.

References

[1] Y. Wang, P. Wang, Y. Bai, Z. Tian, J. Li, X. Shao, L. F. Mustavich, and B. Li, "Assessment of surface water quality via multivariate statistical techniques: A case study of the Songhua River Harbin region, China", *Journal of Hydro-Environment Research*, Vol. 7, pp. 30-40, 2013.

[2] L. J. Wang and Z. H. Zou, "Application of improved attributes recognition method in water quality assessment", *Chinese Journal of Environmental Engineering*, Vol. 2 No. 4, pp. 553-556, 2008.

[3] Y. Liao, J. Xu, and W. Wang, "A method of water quality assessment based on biomonitoring and multiclass support vector machine", *Procedia Environmental Sciences*, Vol. 10, pp. 451-457, 2011.

[4] H. Yoon, Y. Hyun, K. Ha, K. K. Lee, and G. B. Kim, "A method to improve the stability and accuracy of ANN- and SVM-based time series models for long-term groundwater level predictions", *Computers & Geosciences*, Vol. 90, pp. 144-155, 2016.

[5] H. Yoon, S. C. Jun, Y. Hyun, G. O. Bae, and K. K. Lee, "A comparative study of artificial neural networks and support vector machines for predicting ground water levels in a coastal aquifer", *Journal of Hydrology*, Vol. 396, No. 1-2, pp. 128-138, 2011.

[6] P. G. Nieto, J. A. Fernández, V. G. Suárez, C. D. Muñiz, E. G. Gonzalo, and R. M. Bayón, "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", *Applied Mathematics and Computation*, Vol. 260, pp. 170-187, 2015.

[7] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. A. Dujaili, Y. Duan, O. A. Shamma, J. Santamaría, M. A. Fadhel, M. A. Amidie, and L. Farhan, "Review of deep learning : concepts, CNN architectures, challenges, applications, future directions", *Journal of Big Data*, Vol. 8, No. 53, 2021.

[8] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis", *IEEE Access*, Vol. 8, pp. 126385-126400, 2020.

[9] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, and Z. Miao, "Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network", *IFAC-Papers On Line*, Vol. 51, No. 17, pp. 831-836, 2018.

[10] W. Achmad, B. S. Yang, and T. Han, "Combination of independent component analysis and support vector machines for intelligent faults diagnosis of induction motors", *Expert Systems with Applications*, Vol. 32, No. 2, pp. 299-312, 2007.

[11] T. Li, G. Sun, C. Yang, K. Liang, S. Ma, and L. Huang, "Using self-organizing map for coastal water quality classification: Towards a better understanding of patterns and processes", *Science of the Total Environment*, Vol. 628-629, pp. 1446-1459, 2018.

[12] V. Tsvetomil, S. Tsakovski, and V. Simeonov, "Surface water quality assessment using self-organizing maps and Hasse diagram technique", *Chemometrics and Intelligent Laboratory Systems*, Vol. 118, pp. 280-286, 2012.

[13] M. Ladjal, M. Bouamar, M. Djerioui, and Y. Brik, "Performance Evaluation Of ANN And SVM Multiclass Models For Intelligent Water Quality Classification Using Dempster-Shafer Theory", In: *Proc. of International Conference on Electrical and Information Technologies*, 2016.

[14] Y. Li, A. Wright, H. Liu, J. Wang, G. Wang, Y. Wu, and L. Dai, "Land use pattern, irrigation, and fertilization effects of rice-wheat rotation on water quality of ponds by using self-organizing map in agricultural watersheds", *Agriculture, Ecosystems & Environment*, Vol. 272, pp. 155-164, 2019.

[15] S. Mounce, I. Douterelo, R. Sharpe, and J. Boxall, "A bio-hydroinformatics application of self-organizing map neural networks for assessing microbial and physico-chemical water quality in distribution systems", In: *Proc.*

of 10*th International Conference on Hydroinformatics*, 2012.
[16] B. Svetlana and I. Tsiamas, "An ensemble of LSTM neural networks for high-frequency stock market classification", *Journal of Forecasting*, Vol. 38, No. 6, pp. 600-619, 2019.
[17] C. Hu, Y. Duan, S. Liu, Y. Yan, N. Tao, A. Osman, C. I. Castanedo, S. Sfarra, D. Chenf, and C. Zhang, "LSTM-RNN-based defect classification in honeycomb structures using infrared thermography", *Infrared Physics & Technology*, Vol. 102, 103032, 2019.
[18] V. Vapnik, *The Nature of Statistical Learning Theory*, 2000.
[19] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, *The MIT Press*, 2018.
[20] M. H. Bae, T. Wu, and R. Pan, "Mix-ratio sampling: classifying multiclass imbalanced mouse brain images using support vector machine", *Expert Systems with Applications*, Vol. 37, No. 7, pp. 4955-4965, 2010.
[21] M. H. Horng, "Multi-class support vector machine for classification of the ultrasonic images of supraspinatus", *Expert Systems with Applications*, Vol. 36, No. 4, pp. 8124-8133, 2009.
[22] C. Sucheta, L. Vig, and S. Ahmad, "ECG anomaly class identification using LSTM and error profile modeling", *Computers in Biology and Medicine*, Vol. 109, pp. 14-21, 2019.
[23] O. Yildirim, U. B. Baloglu, R. S. Tan, E. J. Ciaccio, and U. R. Acharya, "A new approach for arrhythmia classification using deep coded features and LSTM networks", *Computer Methods and Programs in Biomedicine*, Vol. 176, pp. 121-133, 2019.
[24] M. Toğaçar, B. Ergen, and Z. Cömert, "Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models", *Measurement*, Vol. 153, 107459, 2020.
[25] K. Sourav and S. Ari, "P300 based character recognition using sparse autoencoder with ensemble of SVMs", *Biocybernetics and Biomedical Engineering*, Vol. 39, No. 4, pp. 956-966, 2019.
[26] S. Indu, D. Singh, and A. Khosla, "QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases", *Journal of Advanced Research*, Vol. 4, No. 4, pp. 331-344, 2013.
[27] S. Haidong, J. Hongkai, Z. Huiwei, and W. Fuan, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis", *Mechanical Systems and Signal Processing*, Vol. 95, pp. 187-204, 2017.
[28] A. N. Ahmed, F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, M. Ehteram, and A. Elshafie, "Machine learning methods for better water quality prediction", *Journal of Hydrology*, Vol. 578, 124084, 2019.
[29] M. Djerioui, M. Bouamar, M. Ladjal, and A. Zerguine, "Chlorine Soft Sensor Based on Extreme Learning Machine for Water Quality Monitoring", *Arabian Journal for Science & Engineering*, Vol. 44, No. 3, pp. 2033-2044 2019.