Machine learning techniques in river water quality modelling: a research travelogue

Sakshi Khullar and Nanhey Singh

ABSTRACT

Water is a prime necessity for the survival and sustenance of all living beings. Over the past few years, the water quality of rivers has been adversely affected due to harmful wastes and pollutants. This ever-increasing water pollution is a matter of great concern as it is deteriorating the water quality, making it unfit for any type of use. Contaminated water resources can cause serious effects on humans as well as aquatic life. Hence, water quality monitoring of reservoirs is essential. Recently, water quality modelling using AI techniques has generated a lot of interest and it can be very beneficial in ecological and water resources management. This paper presents the state-of-the-art application of machine learning techniques in forecasting river water quality. It highlights the different key techniques, advantages, disadvantages, and applications with respect to monitoring the river water quality. The review also intends to find the existing challenges and opportunities for future research.

Key words | machine learning, river water quality, water quality evaluation, water quality prediction

Sakshi Khullar (corresponding author)
Guru Gobind Singh Indraprastha University,
M-54, 2nd floor West Patel Nagar, New Delhi,
110008
India
E-mail: *sakshikhullar2809@gmail.com*

Nanhey Singh
HOD & Professor CSE, GGSIPU, AIACTR,
Krishna Nagar Road Chacha Nahru Bal
Chikitsalaya, Geeta Colony, New Delhi, 110031
India

HIGHLIGHTS

- This paper gives a brief literature study, analysis, and comparison of the research done in river water quality evaluation using machine learning models and techniques.
- Finally, it highlights some observations on future research issues, challenges, and needs.

INTRODUCTION

River water is an essential asset for mankind and it is used for various purposes like drinking, traveling, bathing, aquaculture fostering, farming, and energy creation. Hence, a satisfactory degree of water quality is required. The surface water quality in a locale, to a great extent, relies upon the nature and degree of the farming, industrial, and other anthropogenic exercises in the catchments. Maintenance of the water quality becomes indispensable. Predicting the water quality well in advance can greatly help in water quality management of the waterways. To evaluate stream water quality, various chemical parameters like chemical oxygen demand (COD), biochemical oxygen demand (BOD), temperature, dissolved oxygen (DO), pH and conductivity can be used. The traditional approaches for modelling the water quality include parameter-based statistical and deterministic models that need a large amount of information about various hydrological sub-processes to reach the end results. Moreover, these models require precisely determined rate constants/coefficients pertaining to various hydrological, chemical, physical and biological processes, which are largely time and space specific. Many factors influencing the water quality have a complicated non-linear relation with

doi: 10.2166/ws.2020.277

the input variables. Hence, conventional approaches are no longer efficient for solving this problem.

In recent years, several types of research have been conducted on water quality forecast models using machine learning techniques. Various nonlinear models such as artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), support vector machine (SVM) among others have been used for the prediction and forecasting of water resource variables. A huge growth has been seen in machine learning over the past two decades. It is widely used in various commercial areas for practical purposes. Machine learning (ML) includes statistical models and scientific study of algorithms that are effectively used by a computer to perform a particular task without the need for explicit instructions. It can be defined (Jordan & Mitchell 2015) as a subset of artificial intelligence (AI). The base of the ML algorithm is mathematical models that are based on sample data. Today, ML techniques have become the topmost priority for developing software for speech recognition, natural language processing, computer vision etc. The developers of artificial intelligence realize that; it is an easy way to train the systems by presenting the desired required input-output behaviors other than to manually program them to get the required output corresponding to the desired input. Machine learning is becoming popular in industries and computer science areas due to its capability to determine the problems in a complex system, concerned about data issues and the control of logistics chains. A learning problem can be termed as an improvement in the performance through some training or performing some tasks. In machine learning, it can be easily termed as when computers can learn from experience that automatically gets better during execution. If computers are capable of learning from the events and upgrading their performance, their usefulness is automatically increased (Michie *et al.* 1994). Machine learning systems are capable of automatically improving themselves, which makes them more powerful than manual systems. The main aim is to develop a robust system that has self-learning capabilities (Mitchell *et al.* 1990). In the past two decades, machine learning has provided us with self-driving car systems, speech recognition, efficient web search, and a greatly enhanced understanding of the human genome. Machine learning systems are used by us hundreds of times without knowing about it.

MACHINE-LEARNING PREDICTION MODELS WITH RESPECT TO RIVER WATER QUALITY MODELLING

The pollution levels in natural waters like lakes, rivers, streams are increasing day by day. It is one of the worrisome issues faced by humanity. Unclean water can severely affect the health of living beings. So, the management of river water resources is a vital issue to optimize the quality of water. It is very important to predict the quality and pollution level of water based on the statistical studies and available data and also plan prevention solutions to maintain water quality. Machine learning technology is capable of processing the available data, which will help to predict the quality of river water in the future, given certain input parameters. ML can also help in performing detailed study regarding the quality of water in the rivers, estimating pollution levels, and also identifying the main sources of pollution. Eventually, it also helps in planning water quality management programs and developing strategies to combat water pollution. Various types of machine learning algorithms can be used to predict the quality of the water in the river. The following subsection highlights some key machine learning techniques that have been used for water quality prediction and assessment.

Neural networks

The artificial neural network model (Hecht-Nielsen 1992) is an extremely ground-breaking computational procedure for demonstrating complex non-linear relationships between the input variables. The structure of an ANN resembles a human nervous system model, which typically includes three layers: the input layer, where the information is fed into the model and calculation of the weighted aggregate of this information is done; the hidden layer or layers, where information is further processed; and the output layer, where the output values of the ANN are delivered. Each layer comprises one or more fundamental element(s) called a neuron. The data moving through the neuron is modified by weights 'w' and activation functions. The hidden layer sums the weighted inputs and their bias values (b_k) and uses its transfer function to create an output value (y_k). Typical transfer functions are the linear, sigmoid or hyperbolic tangent functions. This process is repeated until the output layer is reached. Figure 1 illustrates a general structure of an ANN having one input layer with n nodes, two hidden layers having m nodes each, and an output layer having two nodes. $w_{n,m}$ denotes the weight between nodes m and n, b_m denotes the bias value of a node m. The number of hidden layers can always increase or decrease depending upon the problem.

Backward propagation neural network (BPNN) is a type of ANN that is defined as 'backward propagation of errors.' It is a technique for training artificial neural networks by altering weights in successive epochs using the concept of back-propagation. This technique also computes the gradient of a loss function with respect to all the weights in the network (Hameed *et al.* 2017).

Hameed *et al.* (Ding *et al.* 2014) predicted the Water Quality Index (WQI) in the Langat and Klang River of Malaysia using DO, BOD, COD, NH3-N (ammoniacal nitrogen), SS (suspended solids), and pH. Two ANN algorithms, namely BPNN and radial basis function neural network (RBFNN) are used to evaluate the WQI. The architecture of RBFNN is similar to the BPNN, except that it employs a radial basis activation function in the hidden layer. BPNN achieves a good performance of coefficient of determination (R^2), root mean square error (RMSE), and Nash–Sutcliffe coefficient (NE) (0.7472, 0.0699, and 0.7701, respectively). However, RBFNN achieves better performance accuracy compared to BPNN with the best R^2 being 0.982. It may be attributed to the fact that the Gaussian radial basis function can mimic the pattern occurrences in unruly disturbances, complex nonlinear systems, and randomness in input water variables that influence the quality index.

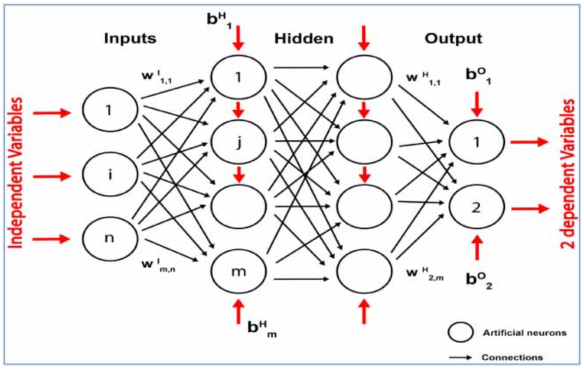

Figure 1 | General ANN structure.

Ding *et al.* (Gao *et al.* 2017) discussed a combination of neural networks, principal component analysis (PCA) and genetic algorithms for predicting river water quality of Taihu Lake, China. The water is classified into two classes; that is, polluted and non-polluted. PCA was effectively used to reduce the size of the input features from 23 to 15. Later on, a Genetic algorithm was used to find the optimal network parameters of the BPNN network. This hybridized BPNN achieves prediction accuracy rates beyond 90 percent as compared to a traditional BPNN, which has an accuracy of around 80 percent only.

Gao *et al.* (Singh *et al.* 2009) estimated eutrophication in Lake Taihu of China using Bayesian regularized BPNN (BRBPNN). The former is used for calculating optimal weights at the input, hidden and outer layers. Eight water quality factors, transparency (SD), pH, water temperature (WT), suspended solids (SS), electrical conductivity (EC), total nitrogen (TN), total phosphorus (TP), and the chlorophyll-a concentration (Chl-a), were used. A model with seven input neurons, six hidden neurons, and one output neuron yielded the best results with R^2 values being 0.77, 0.49, and 0.76 for the training, validation, and test sets. The BRBPNN model performs much better than MLR models and can represent complex relationships between water quality variables and chlorophyll.

Singh *et al.* (Jang 1993) discussed back propagation neural network (BPNN) models to calculate the DO and BOD concentrations in the Gomti River of India. Monthly data of a set of 11 input variables, NH_4-N, K, PO_4, pH, T-Alk, COD, Cl, T-Hard, TS, NO_3-N, and Na, were taken for the past 10 years. ANN for the DO model is composed of one input layer, one hidden layer with 23 nodes and a single output layer with one target outcome, whereas the BOD model is optimized with 11 nodes in the hidden layer. Both models were trained using the Levenberg–Marquardt algorithm (LMA). The RMSE values for DO and BOD are found to be −0.43 and 1.38 respectively for testing datasets. These outputs are capable of capturing long-term trends for DO and BOD.

Adaptive neuro-fuzzy inference system (ANFIS)

ANFIS is actually a multilayer feed-forward neural network that uses fuzzy reasoning to map an input space to an output space (Najah *et al.* 2013). Recently, fuzzy logic systems have been successfully applied to a variety of scientific and engineering problems as a great tool for the modelling of non-linear systems. ANFIS employs a Takagi-Sugeno type fuzzy inference system where every rule is a linear combination of input variables in addition to a constant term. The weighted average of every rule's output is the final output.

Najah *et al.* (Al-Mukhtar & Al-Yaseen 2019) predicted DO concentration at four locations along the Johor river using the ANFIS technique. Water quality data were collected from 1998 to 2007, which included four water quality parameters, ammoniacal nitrogen concentration (NH3-NL), temperature, pH and nitrate (NO_3) concentration. A difficult task in ANFIS is determining the optimal learning parameters (number of membership functions and initial value of step size) before training. Extensive trial and error technique was used to find the best parameters. However, various optimization algorithms could have been utilized. The ANFIS was able to provide satisfactory results with mean absolute prediction error (MAPE) for the four locations being 1.7 for DO1, 1.6 for DO-3, 1.9 for DO-3, and 1.87 for DO-4. The proposed ANFIS model outperforms the MLP model, particularly for extreme values. The capability of the ANFIS model to perform well can be ascribed to the way it identifies input variables. Extreme DO values with specific membership can be easily predicted as fuzzy membership functions representing the input as a range rather than a crisp value.

Al-Mukhtar *et al.* (Specht 1991) used ANFIS to estimate TDS (total dissolved solids) and EC (electrical conductivity) in Abu-Ziriq marsh waters, Iraq. Monthly data of 84 months were collected from 2009 to 2018. After cross-correlation, the input parameters for EC and TDS were found to be total hardness (TH), calcium (Ca^{+2}), sulfate (SO_4), magnesium (Mg^{+2}), and chloride (Cl^{-1}). An additional nitrate (NO_3) parameter was also taken for TDS. ANFIS uses fuzzy Gaussian back-propagation and linear MFs, respectively, for input and output parameters. In comparison with artificial ANNs and multiple regression model (MLR), ANFIS performed better with RMSE (TDS) 169 for calibration and 193.59 for the validation datasets, RMSE (EC) 273.45 for calibration and 246.49 for the validation dataset. ANFIS performs well compared to the other techniques as it integrates the benefit of training ability of neural networks and simplified fuzzy reasoning and rule generation, which eliminates noise and increases accuracy.

Generalized regression neural network (GRNN)

GRNN was designed by D. F. Specht in 1991. It is (Al-Mahasneh *et al.* 2017) a single-pass feed-forward ANN that uses normalized Gaussian kernels in the hidden layer as activation functions. GRNN is composed of input, hidden, summation, division, and output layers. In the training phase, it memorizes every unique pattern and hence does not require any back propagation. After training it with sufficient training sequences, it can achieve a good level of generalization. It has some benefits, which include faster training and high accuracy. On the contrary, one of the disadvantages of GRNN is the progression of the hidden layer size, which can be controlled by using algorithms that contract the growth of the hidden layer by storing only the most significant patterns (Heddam 2014).

Heddam *et al.*(Kumar *et al.* 2004) defined a GRNN-based approach for modelling hourly DO concentrations in the Upper Klamath River, USA. A comparison is made between GRNN and multiple linear regression (MLR) models. The input parameters that are utilized for the two models are the type of water, temperature, electrical conductivity, pH, and sensor depth. A total of six GRNN models were developed depending upon the number of inputs taken. The GRNN models provided a lower mean absolute error (MAE) compared with the MLR models by 0.815 (mg/l), 0.598 (mg/l), and 0.677 (mg/l) for the training, validation, and testing sets, respectively. Hence, GRNN predicted hourly DO concentration more effectively than MLR.

Recurrent neural network (RNN)

A recurrent neural network (Wang *et al.* 2011) is a category of ANNs where relations between nodes form a directed graph along a temporal sequence. They demonstrate a time-based dynamic behavior. Unlike conventional neural networks, RNNs can use their memory or internal state to process an input series.

Wang *et al.* (Antanasijević *et al.* 2013) discussed the application of RNN models to calculate the total phosphorus (TP), total nitrogen (TN) and dissolved oxygen (DO) at three different places in the Gonghu Bay of Lake Taihu during a water diversion period. The input parameters of Elman's RNN were elected by employing principal component analysis (PCA). Simulation results demonstrate that the PCA can be effectively utilized to find the best input parameters for RNN, and this RNN can predict the water quality parameters during the period of water diversion.

Antanasijević *et al.* (Huang *et al.* 2006a) used BPNN, GRNN, and RNN to calculate the DO concentrations in the Danube River. Monthly water quality data was collected over a period from 2004 to 2009. Water quality parameters like water stream, temperature, pH and electrical conductivity were chosen as input features. The input data is divided into training, validation and testing set. ANN model architectures were assessed using mean absolute error and the root mean squared error. RNN gave significantly preferred forecasts of DO over GRNN and BPNN since all predictions for the test data were within the error of less than ±10%. Likewise, an examination of the RNN with the MLR shows that the former exhibits much preferable performance indicators over the latter with its RMSE being 0.59 against 1.49. GRNN gave an excellent performance in the training data but very low performance during testing and validation, which resulted in overfitting. However, even on an unseen test set, both RNN and BPNN provided good generalization. The results are good, but this model should also be checked with other important water quality factors like COD and BOD, which have a deep impact on DO of water.

Extreme learning machine (ELM)

Extreme learning machine was developed by Huang *et al.* (2006a, 2006b, 2012) to improve the learning ability of the standard single hidden layer feed-forward neural networks (SLFN). In ELM, parameters of the hidden nodes like biases and weights are selected randomly and output weights are produced using the least squares method (Huang *et al.* 2006a, 2006b, 2012; Heddam & Kisi 2017). As opposed to the SLFN, in which the weights are determined iteratively using back-propagation, they are randomly initialized and fixed using the ELM. The weights between the hidden and output layers are optimized by solving the Moore–Penrose generalized inverse of matrix. The ELM model may include back-propagation and its speed could be much greater than the conventional SLFNs while retaining good generalization and prediction capability. Figure 2 illustrates a basic schematic structure of an ELM having m input neurons, single hidden layer having n neurons, and one output neuron. A set of $\{x_{1,}x_{2,}\ldots x_{m,}\}$ is given as input features, which are processed to produce a final ouput O_j.

The output of an SLFN with N number of hidden nodes can be represented as (Huang *et al.* 2006a, 2006b, 2012):

$$f(x) = \sum_{i=1}^{N} \beta_i f_i\,(x_{j.}w_i + b_i)$$

where w_i and b_i are the weight and bias components between the input layer and hidden layer respectively. The output weights β are parameters to be estimated. The output function in the hidden layer mapping can be represented as:

$$\mathrm{H}\beta = T$$

$$\mathrm{H} = \begin{bmatrix} f_1(w_1.x_1 + b_1)\ldots. & f_1(w_N.x_1 + b_N) \\ . & . \\ . & . \\ f_1(w_1.x_N + b_1)\ldots. & f_1(w_N.x_N + b_N) \end{bmatrix}$$

$$\beta = H^{+}T$$

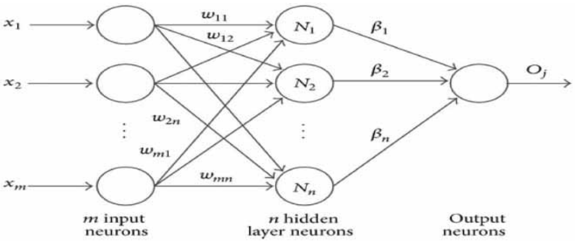

Figure 2 | General ELM structure.

ELM simply solves the above function where H+ is the Moore-Penrose generalized inverse of matrix H

$$\beta = \begin{bmatrix} \beta^T_1 \\ \cdot \\ \cdot \\ \beta^T_N \end{bmatrix} \text{ and T } = \begin{bmatrix} t^N_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ t^T_N \end{bmatrix}$$

Each of the hidden nodes can use a different type of activation function like sigmoid, polynomial, radial basis function, etc.

Heddam *et al.* (Yi *et al.* 2018). utilised a variety of ELM models including simple ELM, ELM with sigmoid activation (S-ELM), ELM with Radial Basis function, online sequential extreme learning machine (OS-ELM), and optimally pruned extreme learning machine (OP-ELM) for predicting the dissolved oxygen in eight river basins in USA. Four water quality variables, pH, temperature, specific conductance and turbidity, were used. Several combinations of these parameters were tried. In some cases, the addition of the turbidity as an input degraded the accuracy of the proposed models. To avoid situations like these, sensitivity analysis should always be carried out first. Out of all the ELM models, OP-ELM had the best accuracy compared to the others at seven stations, with a high coefficient of correlation (R) values (from 0.924 to 0.989) and Nash-Sutcliffe efficiency (NSE) values (from 0.853 to 0.933). The results obtained from different ELM models were found to be more effective when compared and contrasted with multiple linear regression (MLR) and multilayer perceptron (MLP). MLR models had the lowest accuracy after MLP, suggesting that linear models are not so useful in modelling complex relationships of the input parameters with DO.

Hye-Suk Yi *et al.* (Zhu & Heddam 2019) predicted algal blooms in Nakdong River, Korea, using ELM models. Input parameters were collected weekly for a period of three years from 2013 to 2016. The weekly chlorophyll-a concentration was used as a model output, which is the main determining factor of algal blooms. The ELM1 model was used to predict algal blooms and it used solar radiation, total nitrogen, air temperature, N/P ratio, rainfall, total phosphorus and chlorophyll-a concentration as independent parameters. To enhance the prediction accuracy, an ELM2 model was also developed that included an additional upstream chlorophyll-a as an independent parameter. Sigmoid activation function was used and the optimum number of hidden nodes was selected after repeated validation trials. ELM2 model showed off better results at all the weirs with the best RMSE as 6.8 (training) and 13.6 (testing). ELM models were also compared with MLR, ANFIS and BPNN. Results showed that it provides better predictions, lower RMSE, and good generalization as the difference between training and testing error is not large. The model accuracy can be enhanced further by examining water quality in tributaries and incorporating more data in the future.

Heddam *et al.* (Xiang & Jiang 2009) have also applied ELM and MLPNN to predict the DO concentrations of four urban rivers in Three Gorges Reservoir, China, using inputs such as ammonia nitrogen, daily observed water temperature, pH, total nitrogen, total phosphorus, DO, COD permanganate index, and electrical conductivity. Nine different combinations of inputs were used for each of these models. The results indicated that ELM and MLPNN models performed well for Wubu River, acceptable for Yipin River, moderate for Huaxi River, and poor for the tributary of Huaxi River. The best RMSE in the validation phase for MLP is 1.365 and that of ELM is 1.481. ELM and MLP can be used for predicting DO in low pollution rivers but their performance degrades on highly polluted rivers. This may be due to the presence of anthropogenic influences.

Support vector machine (SVM)

Support vector machine was proposed by Vapnik in 1995. Originally, SVM was developed for solving classification problems, but now it is extended to regression-based function estimation as well. It can be practically applicable to continuous, categorical outcomes analogous to Gaussian, logistic, binary and multinomial regression (Huang *et al.* 2006a). The main idea in SVM is to map the training data points into a high dimensional feature space by a kernel function. Figure 3 illustrates a simple SVM structure for binary classification. There exist many decision boundaries called hyperplanes that are separating different classes of data. SVM aims to find an optimal hyperplane that can

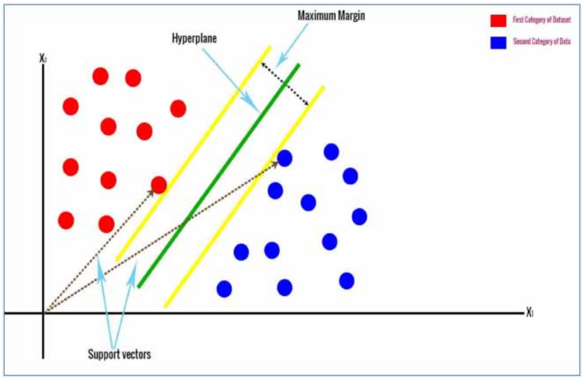

Figure 3 | Simple SVM.

give correct classification and has minimum distance to all the data points. For selecting the best decision boundary, extreme data points closest to the hyperplane are used. These points are called support vectors. The SVM model can be extended to multidimensional scenarios as well.

Consider a training data set $\{(x_{i,}y_{i,},), \mathrm{i} = 1,\ldots, \mathrm{n}\}$, $\mathrm{x} \in \mathrm{Rm}$, $\mathrm{y} \in \mathrm{R}$, where x is the input training data point vector of m parts and y is the corresponding output vector, then the general SVR can be expressed as:

$$f(x) = w.\varphi(x) + b$$

Here, b is the bias, w is the weight vector and $\varphi(x)$ is any non-linear mapping function. The coefficients w and b can be reckoned by minimizing regularized risk function as follows:

$$1/2||w||^2 + C\left(\sum_{i}^{N}(\varepsilon_i + \varepsilon_i^*)\right)$$

Subject to $y_i - w.\varphi(x)$--$b \leq \in + \varepsilon_i, w.\varphi(x) + b - y_i \leq \in + \varepsilon_i^*,\ \varepsilon_i \geq 0,\ \varepsilon_i^* \geq 0$ where C is the regularization parameter, ε_i^* and ε_i are slack variables. The usual technique is to start expressing the problem as a constrained optimization problem, followed by formulating the Lagrangian and then taking the conditions for optimality and finally solving the problem in dual space of Lagrange multipliers. The resulting optimal regression function is:

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_{i,}x) + b$$

Here, $K(x_{i,}x)$ represents the kernel function and α_i, α_i^* are the Lagrange multipliers.

Xiang *et al.* (Jadhav *et al.* 2015) tried predicting water quality using least squares support vector machine (LSSVM) and particle swarm optimization (PSO) in Liuxi River, China. To conquer the weakness of the traditional BP technique, LSSVM joined with PSO is used for estimating the DO and COD values. A total of eight factors such as temperature, alkalinity, chloride, NH_4–N+, NO_2–N, turbidity, pH and hardness were collected over eight years. The parameters σ and γ of the SVM model are tuned by PSO algorithm and a Gaussian kernel is used. Parameters having minimum testing error are found in a shorter time as compared to the Grid Search method. The SVM model outperforms with mean absolute percentage error (MAPE) being 5.8% compared to 6.5% and 6.3% in ARIMA and BPNN models.

Jadhav *et al.* (Mohammadpour *et al.* 2015) predicted the water quality in terms of total faecal coliforms at Gangapur reservoir, Maharashtra, India. They used a combination of least square support vector machines (LS-SVMs) with genetic programming (GP) for data collected over 11 years. GP was used to reduce the number of water quality input variables from 18 to eight. The eight significant parameters were electrical conductivity field (EC_FL), General PH (PH_GEN), DO, total coliforms (Tcol-MPN), total phosphorus (P-Tot), COD, BOD3-27, total alkalinity (ALK-Tot). Later, this data was cross-validated and fed into the SVM model. The parameters of SVM model (gamma and sigma) for radial basis kernel function were found with a trial-and-error method. The best runs were obtained with values of 17 and 7 respectively. The RMSE of LSSVM comes out to be 313.66. In this study, many parameters in GP were fixed after seeing previous studies without analyzing their impact. GP is a technique that does not focus on the insights of the hydrological data and works in a black box mode. Moreover, inappropriately chosen parameters of SVM may result in overfitting or underfitting.

Mohammadpour *et al.* (Haghiabi *et al.* 2018) developed a technique of estimating the water quality index of wetlands in Universiti Sains Malaysia (USM) using SVM and ANNs. A set of 11 water quality variables was reduced to five – pH, COD, DO, AN, and SS – by sensitivity analysis. The SVM model utilized a Gaussian Kernel function with gamma being 0.9. The Grid search algorithm and tenfold cross-validation was used for finding the optimum values of regularization constant and epsilon in SVM. The SVM model was found to be comparable with the conventional BPNN model with $R^2 = 0.9960$ against $R^2 = 0.9988$ in the latter. Both techniques require raw water quality factors that can directly predict WQI. However, the method proposed by the Department of Environment requires a lot of sub-index conversions, which are lengthy and time-consuming.

Amir *et al.* (Haghiabi *et al.* 2018) explored SVM and ANNs for forecasting the water quality of Tireh River, Iran. The input parameters include, bicarbonate, magnesium (Mg^{+2}), temperature (T), pH, total dissolved solids (TDS), calcium (Ca^{+2}), specific conductivity (EC), sulphates, chlorides, sodium (Na^{+}). While formulating these models, it was discovered that tansig transfer function in ANN and RBD kernel function in SVM yielded the best results. However, the SVM technique had more prediction accuracy and reliability compared with ANN.

Decision trees

The decision tree helps to make a decision about a data item. It is one of the most commonly utilized strategies for supervised learning. Decision trees are built through an algorithmic methodology that recognizes approaches that segregate a dataset depending upon various conditions. The objective is to make a model that predicts a target variable by learning basic decision rules inferred from a set of data attributes. Rules are usually in the form of simple if-else constructs. A deep and wide tree with complex rules would usually result in good model fitting. (Areerachakul & Sanguansintukul 2010). It is a form of a tree-based classifier in which internal nodes represent attributes of a dataset, branches correspond to decision rules and leaf nodes denote target outcome.

Srilak *et al.* (Victoriano *et al.* 2019) tried to apply classification and regression tree (CART) for classifying the water quality of different canals in Bangkok. CART is a type of decision tree based upon recursive binary partitioning. It consists of three steps; that is, tree building, tree pruning and optimal tree selection. The water quality class was estimated by collecting data over five years from 2003 to 2007 using six features, which are BOD, ammonia nitrogen (NH_3N), total coliforms (T-Coliform), pH value (pH) and DO. CART exhibits a high accuracy rate of 99.96% in classifying the water quality as compared to MLP, which offers 97.31.

Jayson *et al.* (Bisht *et al.* 2017) predicted the pollution level in Marilao River, Philippines, with the help of a Random Forest decision tree technique. Random Forest works by generating decision trees from a bootstrapped sample taken from several training instances. This procedure is rehashed *n* number of times where '*n*' could be an ideal number of trees produced in the forest. During the development of a tree, a node is split based upon the best random feature set. The presence of randomness results in an expanding bias of the forecast. The outcomes of each produced tree are averaged to give the final prediction. Water quality parameters like DO, BOD, Potential of Hydrogen (pH), and total suspended solids (TSS) were used. Data was gathered from January 2013 up to July 2016 with a total of 473 instances. Random forest model with tenfold cross-validation had an accuracy of 91.75% with a Kappa value of 0.8115, interpreted as 'Strong' in terms of the level of agreement. However, issues like incorporating more training instances, and validating the final model on different datasets need to be addressed.

Bisht *et al.* (Längkvist *et al.* 2014) utilized five types of decision trees such as J48 (C4.5), Random Forest, Random Tree, LMT (logistic model tree) and Hoeffding Tree to classify river water in the Ganga river, India. Monthly data was collected for a period of five years for five water quality parameters like DO, total coliform (TC), temperature, pH, and BOD. All models adopted two types of data division approach; that is, (60–40) % and (80–20) %. Various statistics like Kappa statistics, accuracy, precision, F-measure, mean, absolute error, recall, and root mean squared error were used. Results show that random forest has best classification accuracy of 100% with minimum root mean absolute error (MAE) as 0.006 in (60–40) % and MAE as 0.003 in (80–20) %. The worst performance was of Hoeffding Tree. The models are evaluated in WEKA tool and no information is made available on model parameters like tree size, number of nodes, etc. Model parameters could be changed to increase the effectiveness of other decision tree models.

COMPARATIVE ANALYSIS

Table 1 gives a brief comparison between the above-mentioned techniques and highlights their strengths and weaknesses.

Table 1 | Compartive analysis of different techniques

Author name	Technique	Objective	Water quality parameters	Results	Advantages	Limitation
Hameed *et al.* (2017)	BPNN,RBFNN	Predict WQI in the Langat and Klang River of Malaysia	DO, BOD, COD, NH3-N, SS and pH	RBFNN outperforms BPNN and provides a high accuracy greater than 90 percent for all the combinations	Nonlinear nature of RBFNN makes them highly insusceptible to adversarial noise and model complex relationships between inputs	Trial and error approach was used for finding the optimal number of neurons in all the layers
Ding *et al.* (2014)	PCA, BPNN	To develop a water quality prediction system based on neural networks and genetic algorithms in gives accurate prediction water quality of water quality and offers useful support for real-time early warning systems	pH, NH3-N, volatile phenol, TN, Cr6 + , COD, Mn, TP, BOD5, TCN, COD, petroleum, Cd, Cu, Zn, Pb, Hg, As, Se, F-, sulfide, dissolved oxygen, electrical conductivity, and LAS	This hybridised BPNN achieves prediction accuracy rates beyond 90 percent	PCA reduces the number of significant input factors without information loss which increases the training speed. GA is used to find the optimized network parameters preventing the search process from converging to a local optimum solutions	High model complexity
Gao *et al.* (2017)	Bayesian regularized BPNN	To predict chlorophyll-a concentration Meiliang Bay, Lake Taihu	pH, TN TP, WT SD, SS -EC	BRBPNN model performs much better than MLR models with R^2 0.77, 0.49, and 0.76 for the training, validation, and test sets	Bayesian Regularization improves its generalization ability and optimum number of layers from posterior distribution	However, a full Bayesian approach over all possible weights is computationally very intensive
Singh *et al.* (2009)	BPNN	To predict DO and BOD concentrations in Gomti River, India	NH4-N, PO4, pH, T-Alk, COD, T-Hard, TS, NO_3-N, Cl, K and Na	DO model has RMSE, 1.23 for testing and BOD model as a RMSE of1.38	DO model has better performance than the BOD model	Many DO values are highly deviating from the actual value
Najah *et al.* (2013)	ANFIS	To predict DO concentration at four locations along the Johor river	NH3-NL, temperature, pH and (NO_3)	ANFIS model outperforms MLP. MAPE 1.7 for DO1, 1.6 for DO-3, 1.9 for DO-3, and 1.87 for DO-4	Combines the benefit of neural network with fuzzy logic	ANFIS works well for less number of inputs only, and is difficult to implement on Big Data. Tunable parameters like membership functions require extensive training, which increases complexity
Al-Mukhtar & Al-Yaseen (2019)	ANFIS	To estimate TDS and EC in Abu-Ziriq marsh waters, Iraq	T.H, Ca^{+2} SO_4, Mg^{+2}, Cl^{-1}, NO_3	ANFIS outperforms MLR .RMSE (TDS) 193.59 for validation datasets, RMSE (EC) 246.49 for validation dataset	Good generalization ability	RMSE is quite high. It may be improved by adding input selection or parameter tuning technique
Heddam (2014)	GRNN	To predict dissolved oxygen in the Upper Klamath River, USA	water, pH, temperature, electrical conductivity, and sensor depth	GRNN provided a lower MAE compared with the MLR models by 0.815, 0.598 and 0.677 for the training, validation, and testing sets, respectively	Quick training approach	Growth of the hidden layer size is a bottleneck and careful selection of input parameters is required as RMSE for some GRNN models was high

(*continued*)

Table 1 | continued

Author name	Technique	Objective	Water quality parameters	Results	Advantages	Limitation
Antanasijević *et al.* (2013)	RNN,GRNN	To predict DO concentrations in Danube River	water stream, temperature, pH and electrical conductivity	RNN gave significantly improved forecasts of DO over GRNN and BPNN with RMSE being 0.59	On an unseen new test set, both RNN and BPNN provide good generalization	GRNN gave excellent performance in the training data but very low performance during testing and validation, resulting in overfitting. Water quality factors like COD and BOD which have a deep impact on DO of water should also be considered
Heddam & Kisi (2017)	ELM and its variants(S-ELM, R-ELM, OS-ELM, OP-ELM)	To predict the dissolved oxygen in eight river basins in USA	pH, temperature, specific conductance and turbidity	ELM models outperformed MLP and MLR models with OPELM providing the RMSE values from 0.124 to 0.770	Hidden layers' learning parameters, including the biases and input weights, do not have to be iteratively tuned	ELM is not capable of managing large high dimensional data (Huang *et al.* 2015; Zhang *et al.* 2016) since it needs more hidden nodes compared to the conventional tuning algorithms. Finding the optimal number of hidden layers is the task
Xiang & Jiang (2009)	LSSVM with PSO	To predict COD and DO in Liuxi River, China	Temperature, alkalinity, chloride, NH4 - *N* +, NO2– N, turbidity, pH, hardness	SVM model outperforms ARIMA and BPNN with MAPE being 5.8%	LS-SVM, which solves linear equations instead of a QP problem is computationally less expensive compared to SVM	PSO may easily fall into a local optimum in high-dimensional space
Jadhav *et al.* (2015)	LSSVM with GA	To predict total faecal coliform at Gangapur reservoir, Maharashtra	EC_FL, PH_GEN,DO, Tcol-MPN, P-Tot)\, COD, BOD3-27, ALK-Tot	RMSE of LSSVM comes out to be 313.66	LS-SVM, which is computationally less expensive compared to SVM	Many parameters in GP are fixed after seeing previous studies without analysing their impact. Moreover, inappropriately chosen parameters of SVM may result in overfitting or underfitting
Srilak *et al.*	CART (decision tree)	To classify water quality of canals in Bangkok	BOD), (NH_3N), (T-Coliform), (pH) and (DO)	CART exhibits a high accuracy rate of 99.96% in classifying the water quality compared to MLP	CART algorithm will itself identify the most important variables and reject less influential ones. It can easily handle outliers	May be prone to underfitting if some classes are imbalanced
Bisht *et al.* (2017)	Decision Tree (J48 (C4.5), Random Forest, Random Tree, LMT (logistic model tree) and Hoeffding Tree)	To create a water quality classification model of Ganga river, India	DO, TC, temperature, pH, and BOD	Random forest model outperforms all other models with an accuracy of 100%	Almost all models are capable of achieving a higher classification accuracy above 95% with less error rates	No information is made available on on the model parameters like tree size, no. of nodes etc. Model parameters could be changed to increase the effectiveness of other decision tree models

After reviewing the different types of AI techniques in river quality modelling, it is apparent that one of the most popular models is ANN and its variants. The popularity of ANN can be attributed to the fact that it can model nonlinear, nonstationary, and obscure data without any presumptions between the input and output variables. Novel neural network-based techniques like GRNN, ELM and RNN are also being successfully applied. The study reveals that the performance of neural network-based techniques increases when coupled with a dimensionality reduction technique or a cluster-based technique. Regardless of ANN's success, it has certain limitations like a requirement for large training data for optimum results, overfitting issues, and local minima issues.

ANFIS has shown promising results with nonlinear data and appropriately selected parameters. However, its biggest limitation is the large number of rules that may be created due to too many input variables. Moreover, it requires properly chosen membership functions and other parameters.

SVM is a kernel-based model that is accurate, fast and used for modelling complex relationships. It incorporates structural risk minimization, which leads to good generalization. However, it has high computational complexity and requires carefully chosen Kernel function and parameter values.

Decision trees and its variants are also successfully applied in river quality modelling. They require minimum effort in pre-processing, yet they are more suitable for a classification problem than a regression problem. They have a relatively high time complexity and are expensive to train.

It can be easily stated that the choice of an ML model depends on the problems to be solved, the number and type of input factors, volume and nature of input data, computing resources, acceptable level of prediction accuracy and permissible error rates, as it is unlikely that a single model would outperform others over different datasets and for all types of tasks.

ISSUES, NEEDS AND CHALLENGES

1. Different authors have used different sets of water quality parameters for predicting a particular output. ML models are highly data-driven; that is, their performance greatly depends upon the choice and number of input variables and output variables. The amount of data, location and the time duration has a huge impact on the prediction accuracy. Moreover, sometimes the performance metrics also vary. The models may have low performance when applied to water quality input variables of other rivers. In future, a comparative analysis on this aspect can be done.
2. Several non-measurable factors like climate change, population, sudden wastewater discharges, floods, human construction and socio-economic activities can have a direct impact on the river quality. Hence, one cannot present a definitive conclusion about the applicability of models in a general scope.
3. The performance of a model is directly linked to the number and type of input factors chosen. Selecting the most significant input variables is important as a large number of inputs may cause overtraining and generate local minima, whereas too few inputs may create an incompetent model. To improve the accuracy of water quality prediction, the effect of important biological variables such as dissolved oxygen, pH, conductivity, BOD, COD, and temperature etc. should be considered. Methods like correlation analysis, sensitivity analysis and dimensionality reduction techniques like PCA, CCA, SOM, PFA etc. can be used to find the set of most influential parameters which in turn will help in creating a robust model.
4. The creation of hybrid ML models should be encouraged to enhance their accuracy and performance. Nature-inspired optimizers can be effectively used to address this. Evolutionary optimization techniques might contribute to better performance by improving parameter selection. Some of the popular techniques like GA, ACOR, PSO, FFA and BA have demonstrated better results as compared to the single model.
5. River quality forecasting involves time-series data, which is collected at some regular intervals. Deep learning methods offer a lot of possibilities for time series forecasting, such as automatic learning of temporal dependence and handling time-based structures like trends and seasonality. Deep neural networks and its variants like LSTM, RNN, CNN etc. could be employed to capture

these insights and may result in even better prediction accuracy.

CONCLUSION

This paper gives a brief literature study, analysis, and comparison of the research done in river water quality evaluation using machine learning models and techniques. Finally, it highlights some observations on future research issues, challenges, and needs. The state of the art of ML models in river water quality modelling identified nine groups of ML algorithms; that is, ANN, BPNN, RBFNN, RNN, GRNN, ANFIS, SVM, ELM and Decision Trees with the highest popularity. These models can be successfully used for predicting different water quality factors like dissolved oxygen, Biological Oxygen Demand, Chemical Oxygen Demand, Total Faecal coliforms, pH, Total Dissolved Solids, Conductivity etc. which directly pertain to the wellness of any water system. Such techniques can contribute to the integrated model in different aspects and can provide considerable assistance in laying out strategies for the prevention of water pollution. Most of the published research work has applied only one specific machine learning technique which is largely dependent on the choice of input factors, location and time. There is a need to explore more of integrated hybrid Machine learning that offers better results with improved performance and accuracy. River quality monitoring stations capture data at specific time intervals. The above-mentioned ML strategies can't exploit the temporal sequences or learn from the long-term dependencies of the input variables. Deep learning techniques have proven to be very effective for univariate and multivariate time series forecasting problems. It could be one of the promising directions in establishing a more robust and versatile water quality monitoring system.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Al-Mahasneh, A. J., Anavatti, S. G. & Garratt, M. A. 2017 Altitude identification and intelligent control of a flapping wing micro aerial vehicle using modified generalized regression neural networks. In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 2302–2307.

Al-Mukhtar, M. & Al-Yaseen, F. 2019 Modeling water quality parameters using data-driven models, a case study Abu-Ziriq marsh in south of Iraq. *Hydrology* **6** (1), 24.

Antanasijević, D., Pocajt, V., Povrenović, D., Perić-Grujić, A. & Ristić, M. 2013 Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. *Environmental Science and Pollution Research* **20** (12), 9006–9013.

Areerachakul, S. & Sanguansintukul, S. 2010 Classification and regression trees and MLP neural network to classify water quality of canals in Bangkok, Thailand. *International Journal of Intelligent Computing Research (IJICR)* **1** (1/2), 43–50.

Bisht, A. K., Singh, R., Bhatt, A. & Bhutiani, R. 2017 Development of an automated water quality classification model for the River Ganga. In: *International Conference on Next Generation Computing Technologies*. Springer, Singapore, pp. 190–198.

Ding, Y. R., Cai, Y. J., Sun, P. D. & Chen, B. 2014 The use of combined neural networks and genetic algorithms for prediction of river water quality. *Journal of Applied Research and Technology* **12** (3), 493–499.

Gao, H., Qian, X., Wu, H., Li, H., Pan, H. & Han, C. 2017 Combined effects of submerged macrophytes and aquatic animals on the restoration of a eutrophic water body – a case study of Gonghu Bay, Lake Taihu. *Ecological Engineering* **102**, 15–23.

Haghiabi, A. H., Nasrolahi, A. H. & Parsaie, A. 2018 Water quality prediction using machine learning methods. *Water Quality Research Journal* **53** (1), 3–13.

Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A. & Elshafie, A. 2017 Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Computing and Applications* **28** (1), 893–905.

Hecht-Nielsen, R. 1992 Theory of the backpropagation neural network. In: *Neural Networks for Perception* (H. Wechsler, ed.). Academic Press, George Mason University, Fairfax, VA, pp. 65–93.

Heddam, S. 2014 Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. *Environmental Technology* **35** (13), 1650–1657.

Heddam, S. & Kisi, O. 2017 Extreme learning machines: a new approach for modeling dissolved oxygen (DO) concentration with and without water quality variables as predictors. *Environmental Science and Pollution Research* **24** (20), 16702–16724.

Huang, G. B., Chen, L. & Siew, C. K. 2006a Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks* **17** (4), 879–892. doi:10.1109/TNN. 2006. 875977.

Huang, G. B., Zhu, Q. Y. & Siew, C. K. 2006b Extreme learning machine: theory and applications. *Neurocomputing* **70** (1–3), 489–501. doi:10.1016/j. neucom.2005.12.126.

Huang, G. B., Zhou, H., Ding, X. & Zhang, R. 2012 Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics* **42**, 513–529. doi:10.1109/TSMCB.2011. 2168604.

Huang, G., Huang, G. B., Song, S. & You, K. 2015 Trends in extreme learning machines: a review. *Neural Networks* **61**, 32–48.

Jadhav, M. S., Khare, K. C. & Warke, A. S. 2015 Water quality prediction of Gangapur reservoir (India) using LS-SVM and genetic programming. *Lakes & Reservoirs: Research & Management* **20** (4), 275–284.

Jang, J. S. R. 1993 ANFIS: adaptive-network-based fuzzy inference systems. *IEEE Transactions on Systems, Man, and Cybernetics* **23** (3), 665–685.

Jordan, M. I. & Mitchell, T. M. 2015 Machine learning: trends, perspectives, and prospects. *Science* **349** (6245), 255–260.

Kumar, D. N., Raju, K. S. & Sathish, T. 2004 River flow forecasting using recurrent neural networks. *Water Resources Management* **18** (2), 143–161.

Längkvist, M., Karlsson, L. & Loutfi, A. 2014 A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters* **42**, 11–24.

Michie, D., Spiegelhalter, D. J. & Taylor, C. C. 1994 Machine learning. *Neural and Statistical Classification* **13**, 2–3.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P. & Waibel, A. 1990 Machine learning. *Annual Review of Computer Science* **4** (1), 417–433.

Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., Ab Ghani, A. & Chan, N. W. 2015 Prediction of water quality index in constructed wetlands using support vector machine. *Environmental Science and Pollution Research* **22** (8), 6208–6219.

Najah, A., El-Shafie, A., Karim, O. A. & El-Shafie, A. H. 2013 Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environmental Science and Pollution Research* **21** (3), 1658–1670. doi:10.1007/s11356-013-2048-4.

Singh, K. P., Basant, A., Malik, A. & Jain, G. 2009 Artificial neural network modeling of the river water quality – a case study. *Ecological Modelling* **220** (6), 888–895.

Specht, D. F. 1991 A general regression neural network. *IEEE Transactions on Neural Networks* **2** (6), 568–576.

Victoriano, J. M., Santos, M. L. C. D., Vinluan, A. A. & Carpio, J. T. 2019 Predicting pollution level using random forest: a case study of Marilao River in Bulacan Province, Philippines. *International Journal of Computing Sciences Research* **3** (1), 151–162.

Wang, H., Gao, Y., Xu, Z. & Xu, W. 2011 An recurrent neural network application to forecasting the quality of water diversion in the water source of Lake Taihu. In: *2011 International Conference on Remote Sensing, Environment and Transportation Engineering*. IEEE, pp. 984–988.

Xiang, Y. & Jiang, L. 2009 Water quality prediction using LS-SVM and particle swarm optimization. In: *2009 Second International Workshop on Knowledge Discovery and Data Mining*. IEEE, pp. 900–904.

Yi, H. S., Park, S., An, K. G. & Kwak, K. C. 2018 Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. *International Journal of Environmental Research and Public Health* **15** (10), 2078.

Zhang, J., Feng, L. & Wu, B. 2016 Local extreme learning machine: local classification model for shape feature extraction. *Neural Computing and Applications* **27** (7), 2095–2105.

Zhu, S. & Heddam, S. 2019 Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). *Water Quality Research Journal* **55** (1), 106–118. doi:10.2166/wqrj.2019.053.

First received 29 January 2020; accepted in revised form 9 October 2020. Available online 21 October 2020