Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.26970

218 (2021) 44–51
April

Prediction on water quality of a lake in Chennai, India using machine learning algorithms

D. Venkata Vara Prasad[a], Lokeswari Y. Venkataramana[a], P. Senthil Kumar[b,*], G. Prasannamedha[b], K. Soumya[a], A.J. Poornema[a]

[a]*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai – 603110, India, emails: dvvprasad@ssn.edu.in (D.V.V. Prasad), lokeswariyv@ssn.edu.in (L.Y. Venkataramana), soumya16104@cse.ssn.edu.in (K. Soumya), poornema16075@cse.ssn.edu.in (A.J. Poornema)*
[b]*Department of Chemical Engineering, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai – 603110, India, emails: senthilkumarp@ssn.edu.in (P.S. Kumar), prasannamedhag@ssn.edu.in (G. Prasannamedha)*

Received 16 June 2020; Accepted 22 December 2020

ABSTRACT

The present research work explores different types of machine learning algorithms to estimate the water quality index and the water quality class. The samples were collected from Korattur Lake in Chennai city and tested for its necessary hydro-chemical parameters. The machine learning models such as support vector machine, decision tree, logistic regression, random forest, and naive Bayesian for assessing the water quality with respect to the accuracy and precision of the model. The water quality parameters such as pH, total dissolved salts, turbidity, phosphate, nitrate, iron, chemical oxygen demand, chloride, and sodium are used as a raw dataset. The models are then tested and evaluated to find the best suitable model by comparing and analyzing the accuracy of prediction, the precision rate, and the time taken for execution of all the models. Among all the algorithms employed, the random forest algorithm produces 95% accuracy which is the highest and also consumes the least execution time. From the random forest model, it was found that water quality has 84% of contamination which was attributed to unfit for drinking purpose. Hence, it could be suggested that water quality left disturbed due to anthropogenic activities and improper maintenance.

Keywords: Decision tree; Random forest; Naïve Bayesian classifier; Support vector machine; Classification accuracy; Water quality parameters

1. Introduction

Water is vital for all living organisms. Recently, industrialization and urbanization have led to the scarcity of drinking water by polluting them. Water pollution has become more serious with the rapid development of the economy and urbanization. Water quality plays a major role in public health and the environment because, consumption of impure water leads to many water-borne diseases like cholera, diarrhea, etc. Hence, it is important to check the quality of water before consumption.

Chennai is heavily dependent on the rainfall through which lake water and reservoirs are conserved with natural sources of water, due to the lack of perennial rivers within the city. Pollutants are added to the lakes mainly due to anthropogenic activities like the discharge of sewage, effluents, dumping of solid waste, and release of untreated wastewater. Through this uncontrolled release, lake waters are seriously affected thereby altering the quality of water due to dispersion and dissolution of pollutants. As time prolongs these pollutants are transported

* Corresponding author.

across soil substrate. Further, they are penetrated across soil zones and pollute one of the valuable water resources, groundwater. Groundwater is one of the important sources of drinking water, followed by lake water and reservoirs. The incidence of groundwater pollution is high in areas that are congested and populated heavily as large volumes of waste are concentrated and discharged into natural zones through which hydrochemical parameters are varied. From lakes and reservoirs water is fed for all basic and municipal activities, irrigation, and agriculture [1]. Hence, it is important to know the quality of water before usage. The work focuses on predicting the quality of lake water in Chennai.

Various machine learning models have been built to predict the quality of water till date but the parameters considered in some of the previous works were not sufficient and they couldn't achieve an accuracy of more than 90%. They were not able to handle the multidimensional and imbalanced datasets. Hence, the work involves the machine learning models such as support vector machine (SVM), decision tree (DT), logistic regression (LR), random forest (RF), and naive Bayesian to meet the requirements that the previously used models failed to achieve. The parameters considered are: pH, total dissolved salts (TDS), turbidity, phosphate, nitrate, iron, chemical oxygen demand (COD), chloride, and sodium. These models can handle large datasets that are complex and of nonlinear type. They are well suited for making predictions on time series data and also when the number of parameters considered is large. It also works well with unstructured and semi-structured data. The output obtained is more informative than any other algorithms. Based on these predicted values, the accuracy of the machine learning models are analyzed and compared. The quality of the water in the next 5 y is also known [2–7]. Water quality analysis of Slovenian river using regression tree considered 16 different parameters such as biological oxygen demand (BOD), chlorine concentration (Cl), CO_2 concentration, electrical conductivity, COD ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of ammonia (NH_4), NO_2, NO_3 and dissolved oxygen (O_2), alkalinity (pH), PO_4^{3-}, oxygen saturation, SiO_2, water temperature, and total hardness. The data was obtained from Hydrometeorological Institute of Slovenia from 1990 to 1995 [8].

Muharemi et al. [9] discussed a recently developed water quality prediction system that deals with time-series data. The data was collected from a public water company located in Germany. It incorporated the machine learning and deep learning models such as SVMs, linear discriminant analysis (LDA), logistic regression, artificial neural network (ANN), long short-term memory (LSTM), deep neural network (DNN), and recurrent neural network (RNN). The parameters they chose are as follows: time, turbidity, pH, electrical conductivity, water temperature, chloride (Cl), redox chlorine dioxide, and flow rate. A new machine learning model least squares support vector machine (LS-SVM) that predicts the quality of the Liuxi river in Guangzhou was proposed to overcome the shortcomings of the traditional algorithms, their model combined the LS-SVM with particle swarm optimization (PSO). They considered only the two parameters such as DO and COD. The model was simple to implement and cost-effective that provided solutions within a reasonable time period [10].

The use of naïve Bayesian machine learning algorithms to predict the quality of drinking water could be witnessed in Varalakshmi et al. [11] The model considered parameters such as TDS, pH, nitrate-nitrogen, hardness, and chloride. Thus, their model was designed to assess the quality of water by considering the drinking water standards and calculating the posterior probability. Alternatively, the quality of the Tireh River situated in the southwest of Iran is predicted using the group method of data handling (GMDH), SVM, and ANN. They considered the parameters such as DO, COD, BOD, EC, pH, temperature, K, Na, and Mg [12].

A new methodology was proposed by Ahmed et al. [13] considering the parameters namely, pH, temperature, total dissolved solids (TDS), and turbidity. They used 15 supervised machine learning algorithms such as random forest, multiple linear regression, polynomial regression, gradient boosting algorithm, SVMs, ridge regression, lasso regression, elastic net regression, neural net/multilayer perceptrons (MLP), Gaussian naïve Bayes, logistic regression, stochastic gradient descent, *K* nearest neighbor, decision tree, and bagging classifier to predict the water quality of Rawal Water Lake. The main objective of this work is to predict the quality of water in a region in Chennai using a machine learning algorithm with respect to hydro-chemical parameters. SVM, decision tree, random forest, logistic regression, and naive Bayesian were chosen for the machine learning process. For each algorithm, the program was coded that incorporated water quality index (WQI) as major limits in assessing the quality of water. In order to assess the hydro-chemical nature of lake water, 10 y data was collected. The water samples are collected from the lake every month on a time scale of 10 y. Chosen parameters were studied/calculated based on Bureau of Indian Standards/American Public Health Association (BIS/APHA) standards. These standards helped in evaluating WQI that served as a base limit in coding for predicting the present status for water quality. The study area chosen is Korattur lake which is located north of the Chennai–Arakkonam railway line. It is one of the largest lakes in the western part of the city. It is a chain of three lakes comprising Ambattur Lake, Madhavaram Lake, and Korattur Lake [14]. The machine learning models were used to predict the accuracy of water quality, precision, and execution time.

2. Materials and methods

Fig. 1 depicts the design of the water quality prediction system. The first step is the data pre-processing, which involves cleaning the data and feature selection. Data cleaning is the process of removing missing or inappropriate records from the dataset. Feature selection refers to selecting the most essential features which contribute most to the prediction variable (class). The next step is determining WQI and assigning classes to the data by considering the value of the WQI. The WHO guidelines for drinking water are used for determining the ground truth for drinkable and non-drinkable water. The data is thus classified into binary class (good and bad water samples) and multi-class (excellent, good, average, bad, and poor water samples). Once the classes have been assigned for all the data, the input dataset is divided into training and

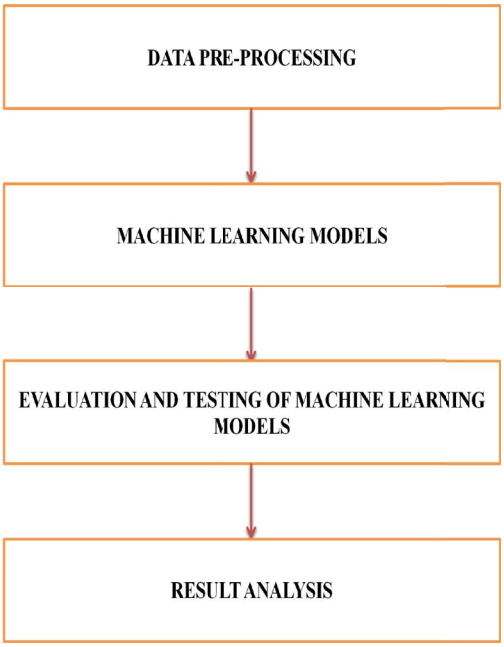

Fig. 1. Water quality prediction system.

testing sets. Among the available data, 80% of the data is used for training and the remaining 20% is used for testing. The models used for training include machine learning models such as decision tree, random forest, SVM, logistic regression, and naive Bayesian. The models are then tested to find the accuracy, precision, and time taken. Based on the outcomes obtained from machine learning models, the results are analyzed to find the best suitable model for our system. The model with the highest accuracy and less execution time is then used for predicting the quality of the water.

2.1. Data pre-processing

Data processing includes four reasonable steps that are to be applied in the prediction. They are dataset collection, dataset description, data splitting, and WQI. All four steps help in water quality assessment of the chosen location that includes water quality parameters as its main tool.

2.1.1. Dataset collection

Korattur Lake is one of the largest lakes in Chennai city. It is spread over 990 acres and is located to the north of Chennai. It has been a major source of drinking water for about 18 y. The dataset for our work was collected from Korattur Lake. The dataset consists of water data for over 10 consecutive years (2010 to 2019). The water samples were collected from the lake over the time period and tested for hydro-chemical quality based on BIS and APHA protocol. Hydro-chemical parameters include pH, TDS, turbidity, phosphate, nitrate, iron, COD, chloride, and sodium. All the collected results were fed in excel worksheets that served as dataset input for the machine learning algorithm. The dataset consists of about 5,000 records that consist of training and testing data. Training data includes standard permissible limits given by WHO/BIS for each parameter whereas testing data includes data collected from tested samples. Nearly 5,000 datasets were incorporated as input in the form of CSV format.

2.1.2. Dataset description

The data set consists of training and testing data of 5,000 records in both binary class as well as multi-class classification consisting of nine parameters as shown in Table 1. The desirable permissible range of parameters for drinking water is shown in Table 2 served as training data.

2.1.3. Data splitting

Before training the deep learning model it is necessary to divide the data into training and testing sets. After splitting the data, the model is trained and tested with certain parts of the data to measure the accuracy of the model's performance. The data was split as a fraction of 4:1 for training and testing respectively. Thus, of the total of 5,000 records, 4,000 samples were used for training, and 1,000 samples for testing.

2.1.4. Water quality index

The WQI is calculated based on the nine parameters such as pH, TDS, turbidity, phosphate, nitrate, iron, COD,

Table 1
Dataset description

Dataset	No. of records	No. of parameters	No. of classes	Class distribution
Korattur Lake	5,000	9	2	Drinkable – 4,325 Non drinkable – 675
	5,000	9	5	Excellent – 649 Very good – 1,831 Good – 1,450 Poor – 620 Very poor – 450

Table 2
Desirable range of drinking water

S. no	Parameters	Suitable range	Reference
1	pH	6.5-8.5	[16–21]
2	Phosphate, mg/L	0.005-0.5	
3	Total dissolved solids, mg/L	300-600	
4	Turbidity, NTU	<5	
5	Nitrate, mg/L	<10	
6	Iron, mg/L	0.3	
7	Chlorides, mg/L	4	
8	Sodium, mg/L	<20	
9	COD, mg/L	3-6	

chloride, and sodium that can provide a simple indicator of water quality. The weights are assigned to each parameter based on the highest difference between minimum value and maximum value of that parameter [15]. After assigning weights for each and every parameter, the quality rating scale is found by using Eqs. (1) and (2):

$$Q_i = \left(\frac{V_i - V_{i0}}{S_i - V_{i0}} \right) \times 100 \quad (1)$$

where V_i stands for the estimated value of nth parameter, S_i is the desirable or permissible range, V_{i0} is the ideal value of nth parameter in pure water. All ideal values are taken as zero for drinking water except pH = 7.0.

Then, the WQI is found by:

$$\text{WQI} = \sum \frac{\left(W_i \times Q_i \right)}{O_{i=1}^{n} W_i} \quad (2)$$

where W_i is the weight allocated to each parameter. Based on the WQI, the classes are classified as shown in Table 3. This WQI value serves as base limits for assessing the quality of samples with respect to each parameters in the dataset.

2.2. Machine learning algorithms

Machine learning algorithms such as naive bayesian, logistic regression, SVM, decision tree, and random forest are used. The parameters considered are: pH, TDS, turbidity, phosphate, nitrate, iron, COD, chloride, and sodium. The output obtained from various algorithms is discussed below.

2.2.1. Support vector machine

SVM differs from other machine learning algorithms. The straight line that is drawn should maximize the distance from the nearest data points of all classes. It ignores the outliers and chooses the maximum distance plane. It is used for classification and regression problems. It is a non-probabilistic linear model but can solve both linear and non-linear problems. It creates a hyperplane that separates the data into classes. It takes the data as input and produces a line that separates the classes as output. There are many ways available for separating the data, but SVM chooses the best optimal decision boundary. It computes the distance between the hyperplanes and support vectors. The goal of SVM is to maximize that distance. The SVM model as discussed by Tong and Koller [22] is shown in Fig. 2.

Table 3
Quality of water based on WQI

Water quality index level	Water quality status	Reference
0–25	Excellent	[15,29]
25–50	Good	
50–75	Poor	
76–100	Very poor	
>100	Unfit for drinking	

2.2.2. Decision tree

Decision tree is a supervised learning algorithm that classifies the data continuously according to certain parameters. To split the records into smaller subsets, it selects the best attribute as a decision node. This process is done recursively to build the tree until no more attributes are left [23]. Fig. 3 shows the model of the decision tree.

2.2.3. Random forest

Random forest is a collection of multiple decision trees which considers the decision of each tree. It divides the dataset into smaller datasets and aggregates the prediction. It samples the data into smaller subsets and trains each of them using a decision tree and aggregates their results. Accuracy increases with increase in number of decision trees [24]. Random forest is shown in Fig. 4.

2.2.4. Logistic regression

Logistic regression is a statistical learning model that makes use of logistic function. It measures the relationship by estimating the probabilities between dependent variables and one or more independent variables. The dependent variables denote the target class and independent variables are attributes used to predict the target class. The logistic function is represented as an *S*-shaped curve. If the value of the input increases above 0, the curve gets closer to 1, and if the value of the input decreases below 0, the curve gets closer to 0. Thus, if the output is more than 0.5, it classifies the outcome as 1, and if it is less than 0.5, it classifies the outcome as 0 [25]. Logistic regression is shown in Fig. 5.

2.2.5. Naive Bayesian

Naive Bayesian is a supervised machine-learning algorithm that is straightforward and can work with millions of records. Naive Bayes classifier follows Bayes' theorem. Bayes classifier computes the probability for each

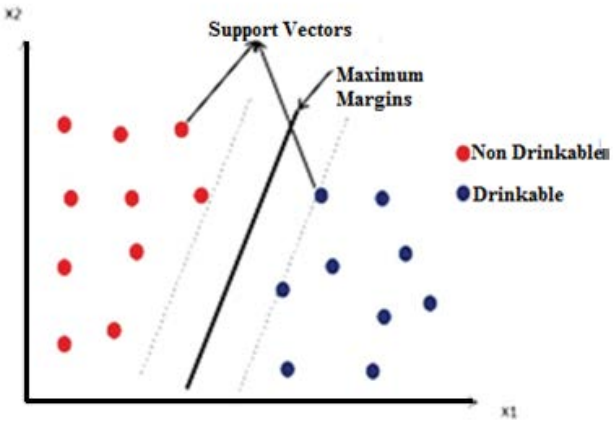

Fig. 2. Support vector machine.

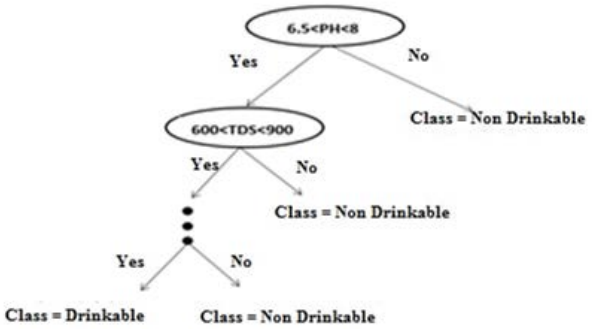

Fig. 3. Decision tree.

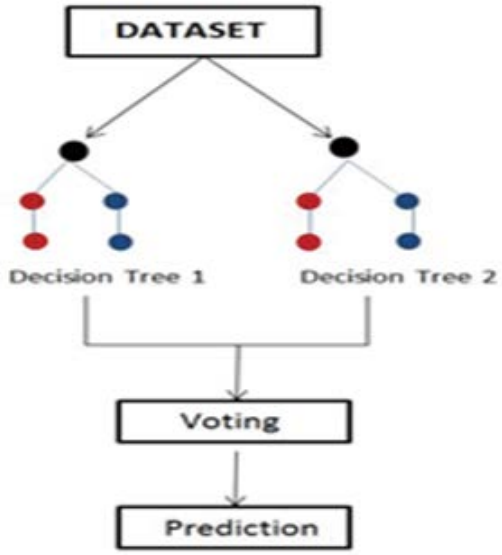

Fig. 4. Random forest.

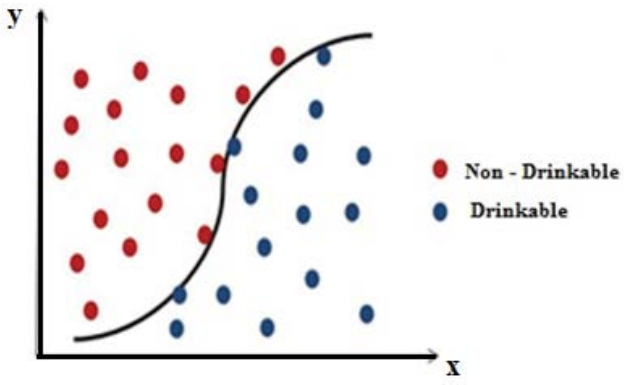

Fig. 5. Logistic regression.

class. Class with highest probability is considered as most likely [26]. Bayes' theorem is given by:

$$P(A \mid B) = P(B \mid A) \cdot P(A) / P(B) \tag{3}$$

where, A, B are the classes A and B; $P(A \mid B)$ is the probability of samples A belonging to class B; $P(B \mid A)$ is the prior probability; $P(A)$, $P(B)$ is the independent probabilities of classes A and B.

3. Results and discussions

3.1. Model evaluation metrics

3.1.1. Confusion matrix

A confusion matrix is a map between the actual and the predicted class label. The confusion matrix is shown in Table 4.

3.1.2. Accuracy

Accuracy is the measure of samples that are correctly predicted in percentage. It is the important metric used to evaluate the performance of the model. It is the ratio of number of correct predictions by total no. of predictions [27]. The formula for accuracy is given by:

$$\frac{(\mathrm{TN}+\mathrm{TP})}{(\mathrm{TN}+\mathrm{TP}+\mathrm{FN}+\mathrm{FP})} \tag{4}$$

3.1.3. Precision

Precision is a useful measure of success of prediction when the classes are very imbalanced. Precision refers to the fraction of the results which are pertinent. That is, precision is the rate of the number of samples correctly predicted as drinkable (class 1) out of all the samples classified as drinkable (class 1) by the model [27,28]. The formula for precision is given by:

$$\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \tag{5}$$

3.1.4. Execution time

The execution time refers to the time taken by the model for training and testing. It is measured in seconds. It includes the time taken for training and testing the model.

3.2. Accuracy of machine learning algorithms

A comparison on the accuracy of the machine learning models is shown in Fig. 6. It is seen that the machine learning algorithms have produced almost the same percentage of accuracy for both binary and multi-class classification. Out of all the algorithms, the random forest algorithm was found to have the highest accuracy of 96% for binary classification and multiclass classification.

3.3. Precision of machine learning algorithms

Precision may also be defined as the number of samples predicted correctly as drinkable water out of all the samples predicted as drinkable. On comparing the calculated precision of all the models as shown in Fig. 7, the binary classification has higher precision than the multiclass classification. Out of all the algorithms, Random forest

Table 4
Confusion matrix

Class names	Class 0 (actual)	Class 1 (actual)	Expansion
Class 0 (predicted)	TN	FP	*True negative (TN)*: samples that are negative, and are predicted to be negative (i.e., non-drinkable water is correctly predicted as non-drinkable). *False positive (FP)*: samples that are negative, but are predicted positive (i.e., non-drinkable water predicted wrongly as drinkable).
Class 1 (predicted)	FN	TP	*True positive (TP)*: samples that are positive, and are predicted to be positive (i.e., drinkable water correctly predicted as drinkable). *False negative (FN)*: samples that are positive, but are predicted negative (i.e., drinkable water is predicted wrongly as non-drinkable).

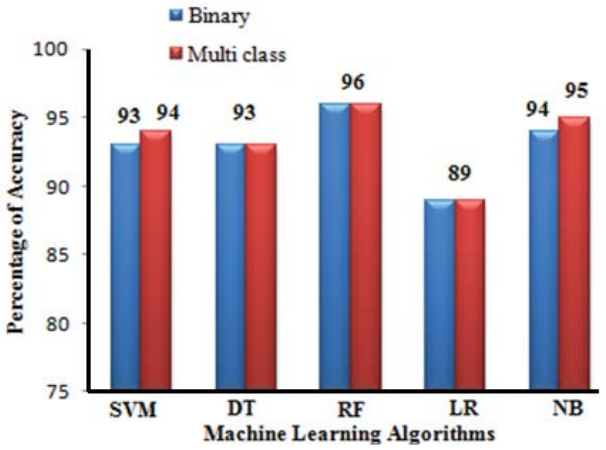

Fig. 6. Accuracy of machine learning models.

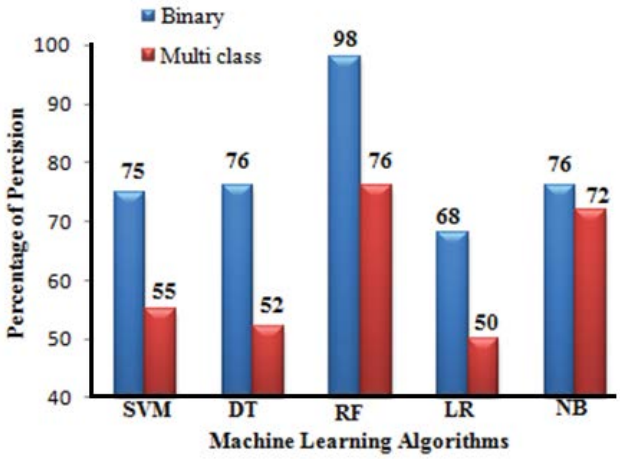

Fig. 7. Precision of machine learning models.

yielded a precision of 98% for binary and 76% for multi-class classification.

3.4. Execution time of machine learning algorithms

The execution time refers to the time taken by the model for training and testing. The execution time of each algorithm is calculated in seconds. The time taken by both the multi-class and binary data is almost the same for all the four models except logistic regression as shown in Table 5. Out of all the algorithms, random forest has the least time of 0.007 s. The analysis and comparison of all the algorithms showed that among the machine learning models, Random Forest resulted with the highest accuracy and precision. The time taken for Random Forest is also the least. So the machine learning model random forest was found to be the best suitable algorithm for our dataset and hence is used for prediction of water quality for the next 3 y. From the random forest algorithm, it could be deduced that in the current location nearly 84% of water could be consumed for drinking and used for domestic purposes and the remaining part of water (16%) has to be treated for further use.

The current work is compared with the work done by Ahmed et al. [13]. They considered four parameters such as pH, turbidity, temperature, and TDS. They used a total of 15 supervised machine learning algorithms such as random forest, multiple linear regression, polynomial regression,

Table 5
Execution time (in seconds) of the machine learning models

Algorithm	Binary classification	Multi-class classification
SVM	4.2 s	4.5 s
Decision tree	0.007 s	0.008 s
Random forest	0.04 s	0.06 s
Logistic regression	0.125 s	3.2 s
Naïve Bayesian	0.007 s	0.01 s

gradient boosting algorithm, SVMs, ridge regression, lasso regression, Gaussian naïve Bayes, logistic regression, neural net/MLP, elastic net regression, stochastic gradient descent, *K* nearest neighbor, decision tree, and bagging classifier to predict the water quality of Rawal Water Lake. Out of all the algorithms used, MLP had the highest accuracy of around 85% and the highest precision of 57%.

Whereas, in our work, we considered the parameters such as pH, TDS, turbidity, phosphate, nitrate, iron, COD, chloride, and sodium. The machine learning algorithms employed include decision tree (DT), random forest (RF), logistic regression (LR), SVM, and naive Bayesian to predict the water quality of Korattur Lake. Out of all the algorithms

employed, the random forest algorithm had the highest accuracy of 96% and the highest precision of around 98%.

Fig. 8 illustrates the comparison of accuracies and precisions. It is seen that the accuracy and precision of our work (random forest) are better than the accuracy and precision of the work done by Ahmed et al. [13]. This is because random forest incorporates many decision trees in decision-making process and it is robust in nature. It does not accept overfitting of incorporated data hence it can be accepted as best machine learning algorithm for the prediction of water quality analysis.

3.5. Prediction of Korattur water quality

Based on the input dataset and coded program the machine learning was performed. It was found from the previous results that random forest gave the most accuracy and precision. Hence further prediction was performed using the random forest learning process. It was found that 84% of the data set were attributed to "not drinkable" as it crossed the permissible limit of prescribed standards and did not fit with WQI index values. Lake water was found to be unfit which may be due to certain anthropogenic activities and improper maintenance.

4. Conclusion

The models used for training and testing include machine learning models such as decision tree (DT), random forest (RF), logistic regression (LR), SVM, and naive Bayesian for binary and multi-class classification. The machine learning models produced an average accuracy of around 93%. Out of the five machine learning algorithms, the random forest was found to be the best suitable algorithm for our work since it produces the highest accuracy of 96%. The random forest also has the highest precision and consumes the least time for execution compared to all the other machine learning algorithms. So, the model forms the basis for predicting the quality of water for the next 3 y. The water quality of the Korattur Lake for the next 3 y (2020–2022) was predicted using RF by collecting the current samples of water from the lake. The attributes of the water were given to the model and the model predicted the quality of water to be "not drinkable". From the prediction, it was found that 84% of water quality is destructed. Based on the data fed as input and WQI value, our prediction suggested that water is unfit for drinking purposes. Lake water found to be unfit for drinking might be due to anthropogenic activities or environmental accidents or calamities like leakage of municipal sewage water into lake, dumping of waste into water, and drainage of industrial outputs into lake water.

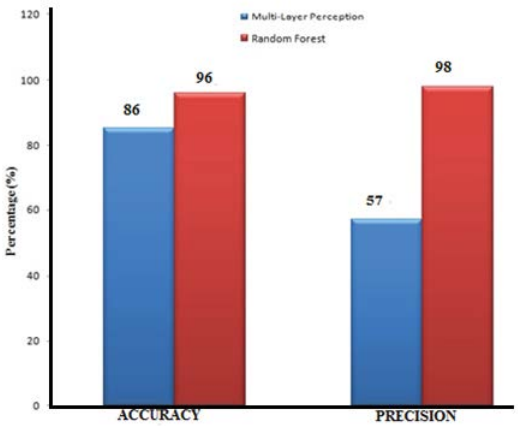

Fig. 8. Comparison of accuracy and precision.

5. Suggestions for prevention of lake water quality

Though Korattur Lake is under maintenance certain preventions could be adopted in order to prevent contamination of the lake which would conserve resources at its own cost. Conservation of resources would help nearby public and municipalities in performing its domestic activities without any interruptions. Precautions like:

- Proper channeling of sewage drains would prevent contamination
- Dumping and disposing of solid waste in dump yard should be encouraged
- Encouraging public in carrying domestic activities nearby lake should be avoided
- Proper conservation of rainwater would help in proper recharge of groundwater and lake water
- Afforestation should be encouraged for conserving water and soil fertility
- Treatment of sewage using novel methodologies may destroy all withheld contaminants
- Discharge of untreated effluents should be prohibited

The work can be extended by including some more classes to the data and the data set can be trained using hybrid models of machine learning and deep learning. Since deep learning algorithms can scale for large data, the hybrid model might be more efficient as it can produce high accuracy as well as can handle large data sets. This is another way of predicting water quality that excludes the interaction of soil with water. It is very well-known that hydrological contaminants get transported across the water and reach seawater that further contaminates marine systems. Hence it is very well-understood that once a part of a water resource is contaminated every part is destroyed. These machine learning models could be easily adopted in predicting the water quality across the world for the proper conservation of resources.

References

[1] A. Kistan, V. Kanchana, A.T. Ansari, Analysis of Ambattur lake water quality with reference to physico-chemical aspects at Chennai, Tamil Nadu, Int. J. Sci. Res., 4 (2013) 944–947.
[2] National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006), Pakistan Council of Research in Water Resources (PCRWR) Islamabad, Islamabad, Pakistan, 2007.
[3] S. Mehmood, A. Ahmad, A. Ahmed, N. Khalid, T. Javed, Drinking water quality in the capital city of Pakistan, Sci. Rep., 2 (2013) 1–6.
[4] A. Azizullah, M.N.K. Khattak, P. Richter, D.-P. Häder, Water pollution in Pakistan and its impact on public health—a review, Environ. Int., 37 (2011) 479–497.

[5] N.M. Gazzaz, M.K. Yusoff, A.Z. Aris, H. Juahir, M.F. Ramli, Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors, Mar. Pollut. Bull., 64 (2012) 2409–2420.
[6] Available at: https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-ravaged-by-raw-sewage-korattur-lake-now-lies-encroached/articleshow/74328391.cms (last accessed May 11, 2020).
[7] Available at: https://timesofindia.indiatimes.com/city/chennai/tamil-nadu-deadline-over-government-seeks-1-month-to-plan-korattur-lake-clean-up/articleshow/74372292.cms (last accessed May 12, 2020).
[8] S. Dzeroski, D. Demsar, J. Grbovic, Predicting chemical parameters of river water quality from bioindicator data, Appl. Intell., 13 (2000) 7–17.
[9] F. Muharemi, D. Logofătu, F. Leon, Machine learning approaches for anomaly detection of water quality on a real-world data set, J. Inf. Telecommun., 3 (2019) 294–307.
[10] Y. Xiang, L. Jiang, Water Quality Prediction Using LS-SVM and Particle Swarm Optimization, Second International Workshop on Knowledge Discovery and Data Mining, IEEE, Moscow, 2009, pp. 900–904.
[11] P. Varalakshmi, S. Vandhana, S. Vishali, Prediction of Water Quality Using Naive Bayesian Algorithm, Eighth International Conference on Advanced Computing, IEEE, Chennai, 2017, pp. 224–229.
[12] H. Haghiabi, A.H. Nasrolahi, A. Parsaie, Water quality prediction using machine learning methods, Water Qual. Res. J., 53 (2018) 3–13.
[13] U. Ahmed, R. Mumtaz, H. Anwar, A.A. Shah, R. Irfan, J. García-Nieto, Efficient water quality prediction using supervised machine learning, Water, 11 (2019) 2210, doi: 10.3390/w11112210.
[14] P. Deepa, R. Raveen, P. Venkatesan, S. Arivoli, T. Samuel, Seasonal variations of physicochemical parameters of Korattur lake, Chennai, Tamil Nadu, India, Int. J. Chem. Stud., 4 (2016) 116–123.
[15] I.N. Balan, M. Shivakumar, P.D.M. Kumar, An assessment of groundwater quality using water quality index in Chennai, Tamil Nadu, India, Chron. Young Sci., 3 (2012) 146–150.
[16] World Health Organization, Guidelines for Drinking-Water Quality, Vol. 1, World Health Organization, Geneva, 1993.
[17] World Health Organization, Water Quality and Health-Review of Turbidity: Information for Regulators and Water Suppliers (No. WHO/FWC/WSH/17.01), World Health Organization, 2017.
[18] A. Colter, R.L. Mahler, Iron in Drinking Water, Pacific Northwest Cooperative Extension, Moscow, Idaho, 2006.
[19] O. Fadiran, S.C. Dlamini, A. Mavuso, A comparative study of the phosphate levels in some surface and groundwater bodies of Swaziland, Bull. Chem. Soc. Ethiop., 22 (2008) 197–206.
[20] Available at: https://water-research.net/index.php/chlorides-and-salts-in-water-future-problem-for-groundwater-users (last accessed May 12, 2020).
[21] D.W. Advisory, Consumer Acceptability Advice and Health Effects Analysis on Sodium, US Environmental Protection Agency Office of Water (4304T), Health and Ecological Criteria Division, Washington, DC, 2003.
[22] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, J. Mach. Learn. Res., 2 (2001) 45–66.
[23] J.R. Quinlan, Decision trees and decision-making, IEEE Trans. Syst. Man Cybern., 20 (1990) 339–346.
[24] A. Liaw, M. Wiener, Classification and Regression by Random Forest, R news, 2/3 (2002) 18–22.
[25] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, New York, NY, 2013.
[26] H. Zhang, The Optimality of Naive Bayes, American Association for Artificial Intelligence, 2004.
[27] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, A. Sattar, B. Kang, Eds., Australasian Joint Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2006, pp. 1015–1021.
[28] E. Goutte, A. Gaussier, Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, D.E. Losada, J.M. Fernández-Luna, Eds., European Conference on Information Retrieval, Springer, Berlin, Heidelberg, 2005, pp. 345–359.
[29] A.K. Chaurasia, H.K. Pandey, S.K. Tiwari, R. Prakash, P. Pandey, A. Ram, Groundwater Quality assessment using water quality index (WQI) in parts of Varanasi District, Uttar Pradesh, India, J. Geol. Soc. India, 92 (2018) 76–82.