International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

REVIEW PAPER ON PREDICTION OF WATER QUALITY PARAMETER USING MACHINE LEARNING

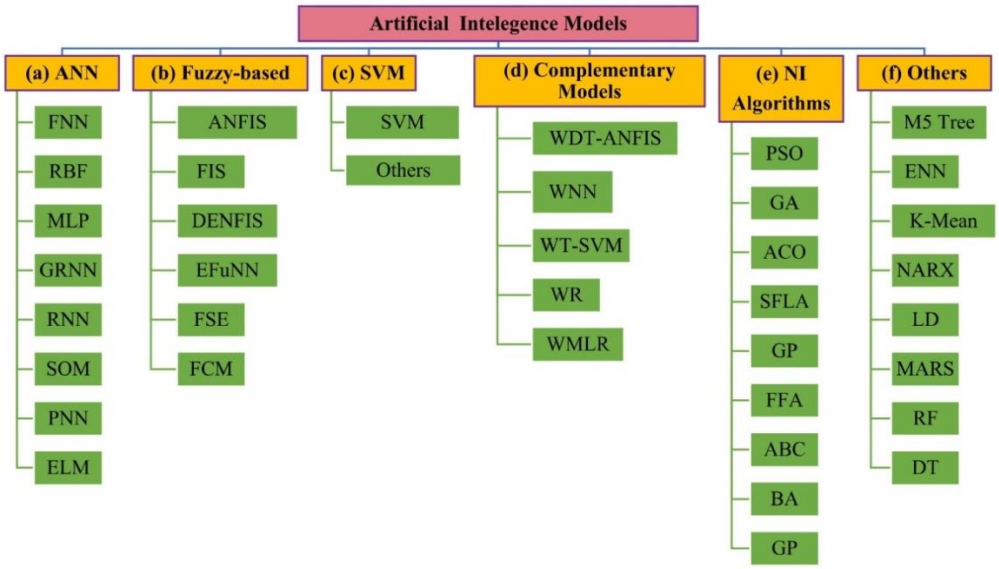
Shubham Shivhare [1], Atul Sharma[2]

[1] Student at Jabalpur Engineering College, enrolled in ME program
[2] Assistant Professor in Jabalpur Engineering College, India

ABSTRACT Surface water pollution become a nuisance for humankind as river water fulfill requirement of a major population and traditional method of water quality assessment and evaluation is inadequate in this era. So using advance method of machine learning in prediction of surface water proves to be helpful to prevent future water accident. As we seen many recent studies of water quality prediction and river water assessment using machine learning approach for better accuracy and less labor and to optimize its overall results. It's become essential to review the recent studies which used Machine Learning algorithms for prediction, analysis, evaluation and assessment of river water quality and different models used in these studies for different environmental conditions. Machine learning models are superior to handle such complex and non linear data such as water quality parameters with greater accuracy, reliability, cost-effectiveness and efficiency as considered as great tool for surface Water Quality monitoring, prediction, future projects and help lawmakers in policy. In this report we reviewed around 17 research papers which uses machine learning approach from different journal and concise it to covers the structure of study, datasets used, methodology analysis, models performance, environment susceptibility, comparative analysis and assessments of Machine Learning models progress in river water quality research. For better management and control of surface water quality and its treatment, this study will help in understanding and analyzing the studies reviewed in this paper and its future application. We can conclude that research on Water Quality prediction using Machine Learning model are inadequate in the context future vulnerability, observing increasing pollution in recent years we require more research in this field. Finally, this study provides breakthrough in Surface Water Engineering and Management to give a new direction to fore coming studies and fortified it scope also gives a comparative approach for its implementation in new studies.

1. **INTRODUCTION** Water is an inorganic, transparent, tasteless, odorless, and nearly colorless chemical substance, which is the main constituent of Earth's hydrosphere and the fluids of all known living organisms. Two major sources of water are surface water and ground water source. From past few decades we have seen river pollution at its peak and water scarcity problem across the world. Increasing sewage and industrial effluent dumped in the river continuously causing heavy damage, even the major river of some part of world get polluted at that extent that the aquatic life in it comes to extinction. Growing concern about the deteriorating condition of river, along with limited funding, world needs the cost effective models and judicious strategies for the management of surface water quality. In developing country with the social and economic development of the country, surface water quality continues to be compromised and deteriorate posing threats to human health. So the requirement of prediction, assessment, and evaluation of river water quality becomes the necessity. This leads focus to method which is easy reliable and efficient and effective for the analysis of water as we have seen the complexity of water related data which cannot be handle with traditional methods. So here the concept of using Machine Learning comes, as many researches had been published in last decade using machine learning algorithm for managing the complexity of large datasets and accurately predict the water quality. Importance of prediction of water quality and its assessment would be beneficial in policy making, river projects and to get early warning of future accident.

Figure 1. Source- Tiyasha et al., 2020



2. REVIEW The summary of the research reported the implementation of Machine Learning and its approaches.

Reference	Input/output	Study Area	Algorithms	Performance metrics	Remarks
Abba et al., 2017	DO ,pH, BOD and WT /DO	Yamuna River, Agra , India	MLR, ANN, ANFIS	DC,RMSE	In this study performance comparison was done between three models, ANFIS with highest accuracy followed by ANN with little variation and both outperformed MLR model for the prediction of DO. Highest correlation shown by pH showing it has maximum effect on the value of DO.
Nouraki et al., 2021	Na, Ca, Mg, Cl, SO_4, K, TH, SAR /TDS, SAR, TH	Karun River, India	MLP, MP5 model tree, SVR, RFR	R^2 , RMSE	This study shows different accurate model for different parameter, RFR predicted TDS, SVR predicted SAR and MLR predicted TH more accurately and these three models had lowest error. PCA method showed that Na, Cl and TH influenced on TDS and Na and Cl influenced on SAR. 20 years data of four station were taken, some data were also missing.
Najah Ahmed et al., 2019	Temperature, EC, salinity, NO3, turbidity, PO4, Cl, K, Na, Mg, Fe Ecoli /AN, SS, pH	Johor River, Malaysia	ANFIS, RBF-ANN, MLP-ANN, WDT-ANFIS	R^2 , CV	In this study shows WDT-ANFIS have the best network architecture, since it outperformed ANFIS and other model. The findings indicate that WDT-ANFIS offered means to improve accuracy and also features the ability to capture temporal patterns in water quality; this enables it to provide meaningful improvements in the generation of forecasts. The model satisfactorily predicted all the water quality parameters.

Bui et al., 2020)	BOD, COD, TS, DO, FC, Ph, PO_4, NO_3, Turbidity, EC/WQI	Talar River, Iran	RF, M5P, RT, and REPT and 12 hybrid algorithm used with BA, CV,CVPS and RFC	R2 , RMSE, MAE, NSE, PBIAS	The models revealed that prediction power is best when the variables with the highest CCs are used. Variables with very low CCs negatively impact predictive power. The level of prediction by the BA-RT model was better than all other models. In order of decreasing performance after BA-RT are RF, bagging-RF, bagging-RT, bagging REPT, RFC-RF, RT, M5P = CVPS-M5P, RFC-M5P, bagging-M5P, REPT, CVPS-REPT, CVPS-RT, RFC-REPT, and RFC-RT. Although the BA-RT hybrid had the highest performance, it didn't predict extreme WQI values accurately.
Csábrági et al., 2017	pH, WT, EC, RF / DO	Danube River, Hungary	MLR, MLP, RBFNN, GRNN	RMSE, MAE, DC, WI	The study using data from the period 1998–2002, found that the nonlinear model performance was better than linear. GRNN and RBFNN outperformed the MLPNN. In order to conduct sensitivity analysis to identify the parameter with the highest influence on the performance of the created models. The sensitivity analysis showed that pH has more influence over DO change than EC, temperature, and runoff. The worst performance was observed in the case of the MLR model even after using with different combinations.
Chen et al., 2020	pH, DO, CODMn, and NH3-N	Songhua, Liaohe , Haihe , Huaihe,	RF, CRF, DCF, DT and CRT LR, LDA, SVM, NB and KNN	precision, recall, F1-score, weighted F1-score,	This study explain that water quality prediction performance of machine learning models may be not only dependent on the models, but also

		Yellow, Yangtze and Pearl River, Taihu , Chaohu, and Dianchi Lake, China			dependent on the parameters in data set chosen for training the learning models. The author adopt a comparitive approach using 10 different models and using big data to improve the performance. Result shows that available big data could improve the performance of both traditional and ensemble learning models in the prediction of surface water quality.
Barzegar et al., 2020	EC, pH, ORP, and WT/DO, chlorophyll-a	Small Prespa Lake, Greece	LSTM, SVR, CNN, DT and Hybrid CNN-LSTM	RSME, MAE, RRMSE, RMAE	The main novelty of this study was to build a coupled CNN–LSTM model to predict water quality variables. Construct DL i.e., LSTM, CNN, and hybrid CNN–LSTM for the first time in the field of water quality modeling which outperformed all standalone models and traditional ML i.e., SVR and DT models to predict DO and Chl-a concentrations. Water quality data were collected at 15-min intervals from June 1, 2012 to May 31, 2013 and datasets trained using the optimal hyper-parameters for each model.
Lu & Ma, 2020	Temperature, DO, pH, Sp. Conductance, Turbidity, FDOM	Tualatin River, USA	RF, XGBOOST, CEEMDAN-RF, CEEMDAN-XGBoost, PSO-SVM, RBFNN, LSSVM, LSTM	RSME, MAPE, RMSPE, U1, U2	This study focus on short term water quality prediction. CEEMDAN used as advanced data de-noising technique. This study collects data from May 1st to July 20th. Author also discussed stability of prediction model and result shows that RF performs best in the prediction followed by XGBoost and these two model outperform other models.
Kadam et al., 2019	pH, EC, TDS, TH, Ca, Mg, Na, K, Cl, HCO3, SO4, NO3	Shivganga River, India	ANM, MLR	R^2, F- test, T-test	In this study, ANN and MLR models are used to find the accuracy of WQI for future prediction of water quality. The

	and PO4/WQI				determination of WQI values are validated through ANN and MLR models. Levenberg–Marquardt three-layer back propagation algorithm was used in ANN architecture. MLR model is used to check the efficiency of ANN prediction. This study found that only single site have water quality of excellent standard according to WQI, however the water is of drinking standard.
Wang et al., 2017	pH, TN, BOD_5, TP, NH3-N, COD, Iron, Copper, Zinc, DO, Volatile Phenol, TDS, Ca, Mg, Na, Cl, HCO_3, SO_4, PO_4, Cr/WQI	Ebinur Lake, China	PSO-SVR	R^2, RMSE, RPD, Solpe, N	This study combines a machine learning algorithm, WQI, and remote sensing spectral indices which are difference index, DI; ratio index, RI; and normalized difference index, NDI through fractional derivatives methods and in turn establishes a model for estimating and assessing the WQI.
Chen et al., 2018	DO, COD, Temperature, pH, BOD, Permaganate index, Ammonia nitrogen, petroleum, volatile phenol.	Yangtze River, China	ABC, BP-NN, IABC, ABC-BP, IABC-BP, PSO-BP	R2, NSE,	This study shows IABC-BP model can be used to forecast water quality and it can increase the forecasting performance of the ABC-BP by searching for the best value of each connection weight and threshold has better network stability, higher learning speed, and stronger approximation ability as compared to the ABC-BP model.
Deng et al., 2015	DO, COD, temperature, EC	Yangtze River, China	ARMA, NAR, RBF-NN, SVM, ANN–GT, OSM	MSE, MAPE, CE, R	This study attempted to use the cloud model theory and fuzzy time series model to handle the uncertain dataset, which extracted the numerical time series into cloud models and represented it by linguistic value (fuzzy sets) by proposed a multi-factor water quality

					time series prediction model based on Heuristic Gaussian cloud transformation, the approximate periodicity of water quality parameter and fuzzy time series model. Fuzzy time series prediction model was applied to generate the computation rule and calculate the predicted value.
Derdour et al., 2022	EC, pH, Na, K, SO4, Mineralization, NO3, Ca, Mg, Cl, HCO3	Naama, Algeria	DT, KNN, DA, SVM, ET,	R2	In this study SVM algorithms classify groundwater quality with high accuracy (95.4%) with standardized data and lower accuracy (88.88%) for raw data. Author found out that SVM is a simple and effective empirical model to simulate water quality, and the method presented in this work is sufficiently general to be applied to a wide range of arid areas. As SVM outperformed all other method.WQI used as indicator.
Ahmed et al., 2019	temperature, turbidity, pH and total dissolved solids	Rawal Lake, Pakistan	Regression Algorithms-LR, PR, RF, GB, SVM, RR, LR, ENR Classification Algorithms-MLP, GNB, LR, SGD, KNN, DT, RF, SVM, GBC, BC	MAE, MSE, RMSE, R2	Author explored an alternative method of machine learning to predict water quality using minimal and easily available water quality parameters also showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently. Hope for future requirement they proposed integrating the findings of this research in a large-scale IoT-based online monitoring system using only the

					sensors of the required parameters.
Shah et al., 2021	Ca,Mg, Na, Cl, SO4, pH, HCO3, TDS, EC	Indus River	GEP, ANN, LRM	NSE, R2 , MAE and RMSE	In this study three algorithm are used and an excellent correlation exhibited among actual and modal stimulated result for both training and testing data. Study showed that performance of the GEP turned out to be the most accurate followed by ANN technique, accuracy of the ANN model decreased on testing data. GEP mathematical expressions for could be easily used in predicting monthly TDS and EC effectively.
Antanasijević et al., 2019	DO	Danube River, Serbia	WNN, SON, PMIS	RMSE, MAE, Bias, MAPE	This study used Location Similarity Index by coupling it with WNN algorithm to prepare a self organizing network based model. Here two groups of monitoring sites were determined, which need two WNN models that have two parallel hidden layers with different activation functions were created. The optimal input combinations were selected using a partial mutual information algorithm, with termination based on the Akaike information criterion. This study concluded that Multiple performance metrics have revealed that the WNN models perform similar or better than multisite DO prediction models published in the literature, while using two to four times less input.
Azad et al., 2017	EC, SAR, TH,	Gorganrood River, Iran	ANFIS, ANFIS-ACOR, ANFIS-DE and ANFIS-GA	R2 , MAPE and RMSE	This study used ANFIS with the application of three evolutionary algorithms including GA, DE and ACOR in performance improvement of

ANFIS for water quality parameters prediction. These algorithms integrated with ANFIS to predict EC, SAR and TH water quality parameters. It was found that ANFIS-DE had this ability to predict different parameters with close performance. Study concluded that meta-heuristic algorithms had significant ability in performance improvement of ANFIS for prediction of river water quality parameters.

3. **CONCLUSION** Prediction of river water quality research interest using machine learning algorithm has grown over from past decades. ML technology has proved to be a powerful tool which has been successfully applied in various field including hydrology and environmental engineering. The review has gone through more than 17 papers which have addressed the river water quality modeling to better assess, predict and manage the current issue of surface water pollution. However, the reviewed literature reveals that these traditional models fail to handle the full aspect of uncertainty, as the datasets available are nonstationary, noisy and nonlinear data when implemented to hybrid model. Every algorithm has their own accuracy making them efficiently working in the prediction of Water Quality modelling and monitoring area. Thus, this review has been limited to the research studies who considered the most common and easily identifiable Water Quality variables. Drawback of the study is constantly changing river Water quality data so, there will always need new models testing them in a new environment. This study provides breakthrough in Surface Water Engineering and Management to give a new direction to fore coming studies and fortified it scope also gives a comparative approach for its implementation in new studies.

4. **FUTURE RESEARCH DIRECTION** There is huge scope in the field of prediction of water quality parameter modelling and these challenges will keep the generation of new innovative ideas. This study helps potential research and challenges which need to be addressed by forthcoming researchers.

Abbreviation Adaptive neuro fuzzy interference system (ANFIS), Alkalinity (AL), Ammonical nitrogen (NH3-N), Artificial neural network (ANN), Artificial bee colony (ABC), Artificial Intelligence (AI), Biochemical oxygen demand (BOD5), bi-carbonate (HCO3), Calcium (Ca), Cadmium (Cd), Calcium sulphate (CaSO4), Carbon di-oxide (CO2), Carbonate hardness (CH), , Decision tree (DT), Determination coefficient/ coefficient correlation (DC), Dissolved oxygen (DO), Dynamic evolving neural fuzzy inference system (DENFIS), Dynamic factor analysis (DFA); Electrical conductivity (EC), Escherichia coli (E.coli), Ensemble neural network (ENN), Extreme gradient boosting (XGB), Gene expression programming (GEP), Gradient boosting (GB), Hardness (H), Machine learning (ML), Magnesium (Mg), Manganese (Mn), Mean Absolute Error (MAE), Mean absolute percentage error (MAPE), Mean percentage error (MPE), Mean square error (MSE), Multi-layer perceptron (MLP), Multi linear regression (MLR), Nash-Sutcliffe efficiency (NSE), Nitrogen (N), Nitrate (NO3), Nitrate Nitrogen (NO3- N), Nitrification Rate (K1), Nitrite (NO2), correlation coefficient (R), Percentage of mean (M%), , Regression model (RM), Regression tree (RT), Root mean square error (RMSE), Support vector classification (SVC), Support vector machine (SVM), Support vector regression (SVR); Total dissolved solid (TDS), Total nitrogen (N Tot.), Total coliform (TC), Total organic carbon (TOC), Total phosphorus (PTot.), Total solid (TS), Water quality (WQ), Water quality index (WQI), Water temperature (WT), Wavelet neural network (WNN).

REFERENCE:

1. Tiyasha, Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology*, *585*, 124670. https://doi.org/10.1016/j.jhydrol.2020.124670
2. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, *11*(11), 2210. https://doi.org/10.3390/w11112210
3. Derdour, A., Jodar-Abellan, A., Pardo, M. Á., Ghoneim, S. S. M., & Hussein, E. E. (2022). Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water*, *14*(18), 2801. https://doi.org/10.3390/w14182801
4. Antanasijević, D., Pocajt, V., Perić-Grujić, A., & Ristić, M. (2019). Multilevel split of high-dimensional water quality data using artificial neural networks for the prediction of dissolved oxygen in the Danube River. *Neural Computing and Applications*, *32*(8), 3957–3966. https://doi.org/10.1007/s00521-019-04079-y
5. Abba, S. I., Hadi, S. J., & Abdullahi, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science*, *120*, 75–82. https://doi.org/10.1016/j.procs.2017.11.212
6. Shah, M. I., Alaloul, W. S., Alqahtani, A., Aldrees, A., Musarat, M. A., & Javed, M. F. (2021). Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models. *Sustainability*, *13*(14), 7515. https://doi.org/10.3390/su13147515
7. Chen, S., Fang, G., Huang, X., & Zhang, Y. (2018). Water Quality Prediction Model of a Water Diversion Project Based on the Improved Artificial Bee Colony–Backpropagation Neural Network. *Water*, *10*(6), 806. https://doi.org/10.3390/w10060806
8. Csábrági, A., Molnár, S., Tanos, P., & Kovács, J. (2017). Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. *Ecological Engineering*, *100*, 63–72. https://doi.org/10.1016/j.ecoleng.2016.12.027
9. Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H., & Sayyahi, F. (2017). Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River). *KSCE Journal of Civil Engineering*, *22*(7), 2206–2213. https://doi.org/10.1007/s12205-017-1703-6

10. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, *171*, 115454. https://doi.org/10.1016/j.watres.2019.115454
11. Wang, X., Zhang, F., & Ding, J. (2017). Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12853-y
12. Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, *249*, 126169. https://doi.org/10.1016/j.chemosphere.2020.126169
13. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612
14. Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084
15. Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N., & Sankhua, R. N. (2019). Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. *Modeling Earth Systems and Environment*, *5*(3), 951–962. https://doi.org/10.1007/s40808-019-00581-3
16. Nouraki, A., Alavi, M., Golabi, M., & Albaji, M. (2021). Prediction of water quality parameters using machine learning models: a case study of the Karun River, Iran. *Environmental Science and Pollution Research*, *28*(40), 57060–57072. https://doi.org/10.1007/s11356-021-14560-8
17. Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*, *34*(2), 415–433. https://doi.org/10.1007/s00477-020-01776-2
18. Deng, W., Wang, G., & Zhang, X. (2015). A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting. *Chemometrics and Intelligent Laboratory Systems*, *149*, 39–49. https://doi.org/10.1016/j.chemolab.2015.09.017