Water Status Prediction Using Machine Learning Algorithm

Indra.S[1], Mr.A.ArulAmalraj[2]
[1]Dept of MCA
[2]Assistant Professor, Dept of MCA
[1, 2] Francis Xavier Engineering College, Vannarpettai

Abstract- *The major goal of this project is to use machine learning techniques to measure water quality. A portability is a numerical phrase that is used to assess the quality of a body of water. The following water quality parameters were utilized to assess the overall water quality in terms of portability in this study.ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbiditywere the parameters. To depict the water quality, these parameters are used as a feature vector. To estimate the water quality class, the paper used two types of classification algorithms: Decision Tree(DT) and K- Nearest Neighbor (KNN). Experiments were carried out utilizing a real dataset containing information from various locations around Andhra Pradesh, as well as a synthetic dataset generated at random using parameters. Based on the results of two different types of classifiers, it was discovered that the KNN classifier outperforms other classifiers. According to the findings, machine learning approaches are capable of accurately predicting the portability. Portability, Water Quality Parameters, Data Mining, and Classification are all index terms.*

Keywords- Machine Learning Supervised Learning, K-Nearest Neighbour (KNN), Decision Tree, Hyper.Parameter Tuning, Python Programming.

I. INTRODUCTION

The various components that impact water quality make it a complicated issue. The different uses of water are closely tied to this idea. Various needs need different solutions standards. Water quality prediction is the subject of extensive research. The physical and chemical properties of water are typically assessed in relation to the intended use of the water. The values for each variable must then be defined as acceptable and unacceptable. Water is deemed suitable for a given use when it complies with the pre-set requirements. The water has to be filtered before use if it does not meet these requirements. Many physical and chemical characteristics can be used to evaluate the quality of water.

As a result, it is not practical to adequately define water quality on a spatial or temporal basis by observing the behavior of each individual variable alone. Combining the values of several physical and chemical factors into one value is the trickier approach. Each variable index had a quality value function (often linear) that represented the equivalency between the variable and its quality level. These calculations have been performed using physical parameter values obtained from water sample studies or direct measurements of a substance concentration. Evaluating how machine learning algorithms might be used to forecast water quality is the main objective of this study. The environment and the entire public health are directly impacted by water quality. Water is utilized for many purposes, including drinking, farming, and industrial. In recent years, the growth of water games and other activities has significantly aided in drawing tourists. Due to their accessibility, rivers have been utilized more frequently for the growth of human societies than other water sources. Other water sources, such groundwater and ocean, can occasionally help with issues. According to Ener et al.(2017), measurements of dissolved oxygen (DO), chemical oxygen demand (COD), biochemical oxygen demand (BOD), electrical conductivity (EC), pH, temperature, K, Na, and Mg, among other water quality factors, have been recommended. Governments have built hydrometrystations along rivers that flow through urban areas, agro-industrial projects, industrial estates, and rivers that connect dam reservoirs to achieve this goal. The components of water quality are measured and the stage-discharge relation is defined in hydrometry stations. Values obtained from hydrometry stations offer the fundamental data needed for feasibility studies and the creation of water conservation initiatives. In order to determine the kind of irrigation system.

II. RELATED WORKS

PCRWR. National Water Quality Monitoring Programme et al [1], Fifth Monitoring Report (2005–2006); Pakistan Council of Research in Water Resources Islamabad: Islamabad, Pakistan, 2007.The National Water Quality Monitoring Programmed (NWQMP) was initiated by Pakistan Council of Research in Water Resources (PCRWR) in 2002. It was the premier project of the year which generated the first detailed water quality profile of 23 major cities of the country. The NWQMP continued for five years (2002-2006). This report is the final and fifth technical report of 2005-06 and presents the results of the final phase of the monitoring program.

Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju et al [2], M. Development of a water quality index (WQI) for the Loktak Lake in India. Appl. Water Sci. 2017, 7,2907–2918.The present work was carried out to assess a water quality index (WQI) of the Loktak Lake, an important wetland which has been under pressure due to the increasing anthropogenic activities. Physicochemical parameters like temperature (Tem), potential hydrogen (pH), electrical conductivity (EC), turbidity (T), dissolved oxygen (DO), total hardness (TH), calcium (Ca), chloride (Cl), fluoride (F), sulphate (SO2 4), magnesium (Mg), phosphate (PO3 4), sodium (Na), potassium (K), nitrite (NO2), nitrate (NO3), total dissolved solids (TDS),total carbon (TC), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) were analyzed using standard procedures.

Thukral, A.; Bhardwaj, R.; Kaur, R. et al [3], Water quality indices. Sat 2005, 1, 99.The quality of water may be assessed using the Water Quality Index (WQI) developed by National Sanitation Foundation, USA in 1970. The index takes into account nine water quality parameters: Dissolved oxygen, Faecal coliform, pH, BOD5, Temperature difference (1 mile), Total phosphate, Nitrate, Turbidity and Total solids. Water is graded for its quality from the worst to best on a scale of 0 to 100. The quality of water for a given parameter is represented by ‘Q’ value. Each water quality parameter is assigned a value known as Weighting factor.

Srivastava, G.; Kumar, P. et al [4], Water quality index with missing parameters. Int. J. Res. Eng. Technol. 2013, 2, 609–614.In this paper a formula will be found to calculate water quality index when the numerical value of some of it’s quality parameters are missing.The standard formula to calculate water quality index has nine water quality parameters- biochemical oxygen demand,dissolved oxygen, pH, nitrate, phosphate, faecal coliform, turbidity, total dissolve solids and temperature.

The Environmental and Protection Agency, “Parameters of water quality,” Environ. Prot., p.133, 2001 et al [5], The Upper Chongwe River Catchment has recently been overexploited forwater resources with increased complaints by various water users about the deteriorating quality of surface water within the sub-catchment.This study was motivated by the need to be investigated.

P. Zeilhofer, L. V. A. C. Zeilhofer, E. L. Hardoim, Z. M. . Lima, and C. S. Oliveira, et al [6],“GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil,” Cadernos de SaúdePública, vol. 23, no. 4, pp. 875–884, 2007.A cross-sectional study utilizing spatial analysis techniques was conducted to study water quality problems and risk of waterborne enteric diseases in a lower-middle-class urban district of Cuiabá, the capital of Mato Grosso State, Brazil.

M. A. Kahlown, M. A. Tahir, and H. Rasheed et al [7], National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006) Pakistan Council of Research in Water Resources Islamabad, Islamabad, Pakistan,2007.To explore the possibility of fluoride toxicity, 747 water samples were collected from surface water and groundwater sources of 16 major cities of Pakistan, adopting a uniform sampling design with distribution of samples: Mirpur Khas (55), Peshawar (38), Risalpur (35), Quetta (81), Ziarat (21),Loralai (21), and Mastung (37). Lahore (79), Kasur (46), Faisalabad (30), Khushab (50), Chakwal (51), Mianwali (30), Jhelum (53), Bahawalpur (60), Karachi (60),

K. Farrell-Poe, W. Payne, and R. Emanuel et al [8], Water Quality & Monitoring, University of Arizona Repository, 2000, http://hdl.handle.net/10150/146901.During the last years, water quality has been threatened by various pollutants. Therefore, modeling and predicting water quality have become very important in controlling water pollution. In this work, advanced artificial intelligence (AI) algorithms are developed to predict water quality index (WQI) and water quality classification (WQC).

D. F. Hayes, J. W. Labadie, T. G. Sanders, and J. K. Brown et al [9], “Enhancing water quality in hydropower system operations,” Water Resources Research, vol. 34, no. 3, pp. 471–483, 1998.The quality of impounded waters often degrades over time because of thermal stratification, sediment oxygen demands, and accumulation of pollutants. Consequently, reservoir releases impact water quality in tailwaters, channels, and other downstream water bodies.

J. Liu, C. Yu, Z. Hu et al [10], “Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network,” IEEE Access, vol. 8, pp. 24784–24798, 2020.In the smart mariculture, the timely and accurate predictions of water quality can help farmers take countermeasures before the ecological environment deteriorates seriously. However, the openness of the mariculture environment makes the variation of water quality nonlinear, dynamic and complex.

III. THEORY

We are going to implement a water quality prediction using machine learning techniques. In this technique, our model predicts that the water is safe to drink or not using some parameters like Ph value,conductivity, hardness, etc. PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. Pure water is not a good conductor of electric current rather it is a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.

The proposed system is intended to determine potability. It is divided into two phases, one for training and the other for testing. The following procedures are carried out in both sections.Data on training pH and hardness testing data Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe something. The data set was chosen as follows: The collection of essential parameters that affect water quality, identification of the number of data samples, and definition of the class labels for each data sample present in the data are all factors that go into selecting the water quality data set, which is a prerequisite to model construction. Ten parameters make up the data sets used in this study. pH value and hardness are examples of these factors. Solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe the properties of a substance.

A 1. Research Methodology

The proposed approach, however, is not constrained by the number of parameters or the selection of parameters. A k-fold cross validation technique is employed to set the learning and testing framework in this study, corresponding to each data sample in the data set. The dataset is separated into k-disjointed sets of equal size, each with roughly the same class distribution, using this technique. This division subsets are utilised as the test set in turn, with the remaining subsets serving as the training set. These are Decision Tree (DT) and K-Nearest Neighbour (KNN) methods. In terms of the underlying relational structure between the indicator parametersand the class label, each of these strategies takes a different approach. As a result, each technique performance for the same data set is likely to differ. Validating the performance of different classifiers on an unknown data set: Data mining provides several metrics for validating the performance of different classifiers on an unknown data set. A repeated cross-validation procedure in the Matlab caret package was used to create the learning and testing environment.

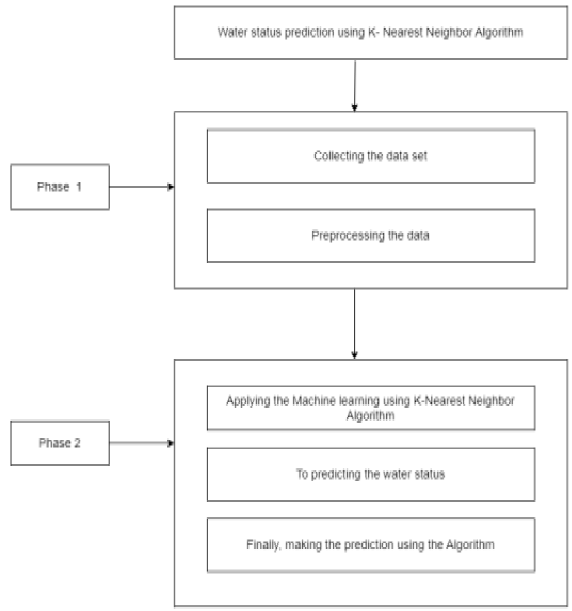


Figure 1. Research Methodology

A 2. Algorithm Implementation

The k-nearest neighbors (KNN) algorithm is a straightforward, easy-to-implement AI calculation that can be used to address both grouping and relapse issues. The K-NN method saves all available data and arranges another data point based on proximity. This means that as new knowledge becomes accessible, it can be quickly grouped into a useful suite classification using K-NN calculations. The K-NN equation can be used for both Regression and Classification, but it is most commonly used for Classification problems. K-NN is a non-parametric measure, which means it makes no assumptions about simple data.To estimate river water quality class, two data mining methods were used: Decision Tree(DT) and K- Nearest Neighbour(KNN). These methods are both parametric and nonparametric classifiers, and their goal is to develop a function that maps input variables to output variables from a training dataset. Because the function's form is unknown, different algorithms make different assumptions about the function's form and how training data is learned to produce the output. The parametric learning classifier makes more confident assumptions about the data. If the assumptions for any data set are true, these classifiers will make rectification judgments. However, if the assumptions are incorrect, the same classifier performs poorly. In order to learn classification tasks, these classifiers do not rely on the quantity of the sample data set; rather, their working principles are their assumptions. This classifier is susceptible to prediction mistakes such as bias, in addition to its parametric character. When the model makes multiple assumptions, the Decision

Tree yields substantial bias. Nonparametric classifiers, unlike parametric learning classifiers, do not make any assumptions about the form of the mapping function, and by not making any assumptions, they are having more accuracy. These classifiers can create any function from the training data set.

IV. EXPERIMENTS AND RESULTS

A 1. Simulation Environment

The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience. It is a fully open-source product, and users can use every functionality available for free. It supports more than 40 languages including Python, R, and Scala. A notebook is a mutable file saved in ipynb format. Jupyter notebook has a notebook dashboard to help users manage different notebooks.

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality. The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.Jupyter Notebook allows users to compile all aspects of a data project in one place making it easier to show the entire process of a project to your intended audience. Through the web-based application, users can create data visualizations and other components of a project to share with others via the platform.

A 2. Architecture diagram

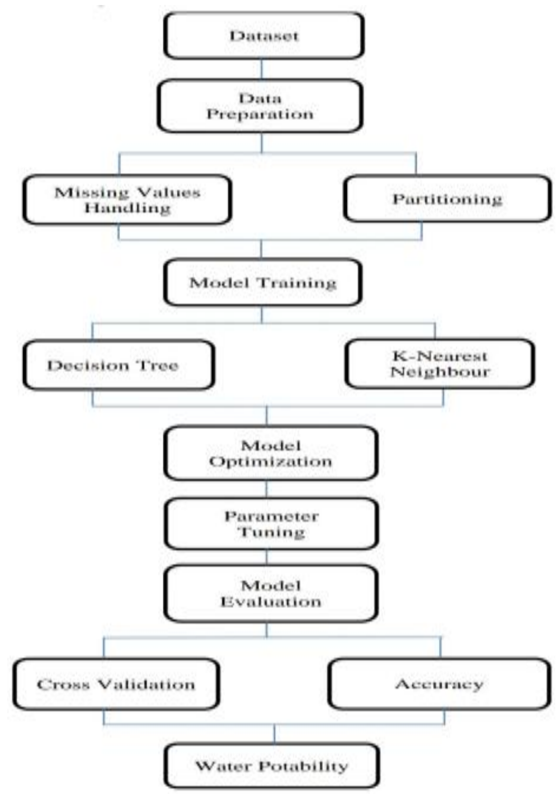


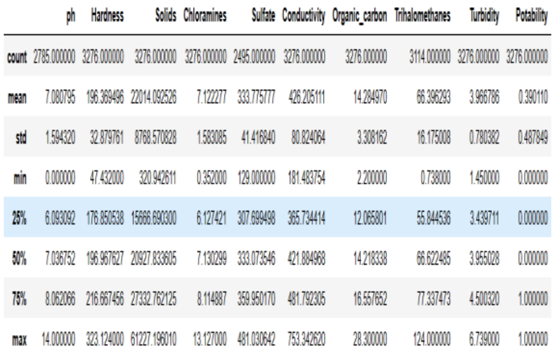
Figure 2. *Architecture diagram*

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890456	20791.31898	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.05786	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.54173	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771	0
4	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Figure 3. Exploratory Data Analysis

```
ph                 491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

Figure 4. Handle the missing values



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690300	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833605	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762125	8.114887	359.950170	481.792305	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196010	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Figure 5.Visualize the parameters of values

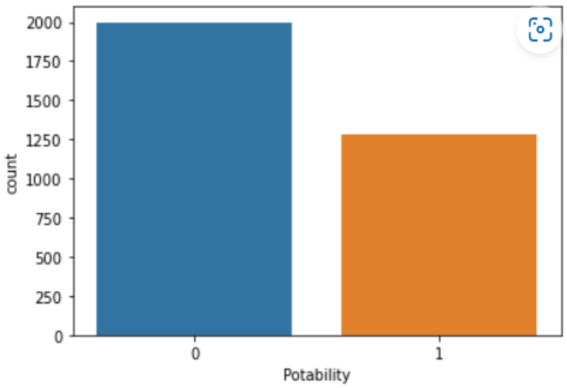


Figure 6. Visualize the portability

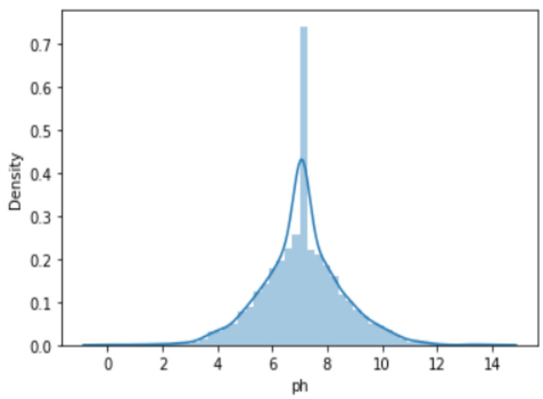


Figure 7. Visualize the PH value

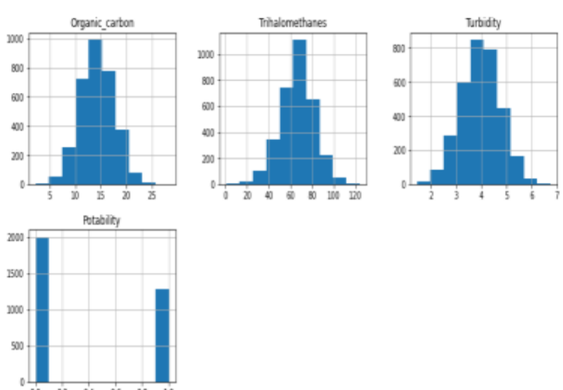


Figure 8. Visualize all the features of dataset

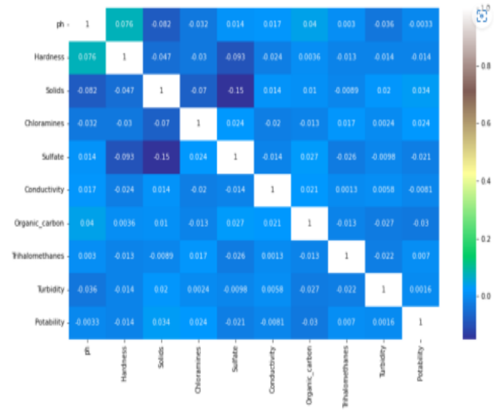


Figure 9. Visualize the correlation

A 3. Performance Metrics

Serial No.	Parameters	Standard
1	Ph	6.5-8.5
2	Dissolved Oxygen	5.0-6.0mg/l
3	Chemical Oxygen Demand	250mg/l
4	Total dissolved solids	500mg/l
5	Chlorides	250mg/l
6	Sulphate	200mg/l
7	Nitrate-nitrogen	45mg/l
8	Fluoride	0.6-1.2mg/l
9	Phenol	0.001mg/l
10	Iron	0.3mg/l
11	Conductivity	0.5-1.5mS/cm
12	Magnesium	30mg/l
13	Calcium	75mg/l

TABLE I. algorithms prediction result

V. DISSCUSSION AND CONCLUSION

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

VI. FUTURE SCOPE

The proposed approach has a Humans cannot drink saline water, but, saline water can be made into freshwater, for which there are many uses. The process is called "desalination", and it is being used more and more around the world to provide people with needed freshwater. One technology designed to help produce more freshwater is desalination plants. Water desalination is the process of removing salt from seawater to produce fresh water that can be processed further and safely used. A desalination plant converts about half of the water it receives into drinkable water. By using this prediction we can easily predict saline water can made into fresh water as well as we can predict the salt water for the laboratory purpose to purify the human lungs etc.

REFERENCES

[1] L. Metcalf and H. P. Eddy, Wastewater Engineering: Treatment and Resource Recovery, 5th Edition. McGraw Hill, Chapter 3, 2014.

[2] A. H. Zare, "Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters," Journal of Environmental Health Science & Engineering, vol. 12, no. 1, pp. 1-8, Jan. 2014.

[3] L. Li, P. Jiang, H. Guang, L. Dong, G. Wu, and H. Wu " Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China," Environmental Science and Pollution Research, vol. 26, no. 19, pp. 19879-19896, Mar. 2019.

[4] J.-S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference Systems," IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May. 1993.

[5] A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring," Environmental Science and Pollution Research, vol. 21, no. 3, pp. 1658 1670, Aug. 2013.

[6] J. Wan, M. Huang, Y. Ma, W. Guo, Y. Wang, H. Zhang, W. Li, and X. Sun, "Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system," Applied Soft Computing, vol. 11, no. 3, pp. 3238-3246, Apr. 2011.

[7] W. Deng, G. Wang, and X. Zhang, "A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting," Chemometrics and Intelligent Laboratory Systems, vol. 149, pp. 39-49, Dec. 2015.

[8] P. Singh, "Rainfall and financial forecasting using fuzzy time series and neural networks based model," International Journal of Machine Learning and Cybernetics,vol. 9, no. 3, pp. 491- 506, May. 2018.

[9] A. K. Kadam, V. M. Wagh, A. A. Muley, B. N. Umrikar, and R. N. Sankhua, "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, vol. 5, no. 3, pp. 951-96, Mar. 2019.

[10] A. Sarkar and P. Pandey, "River Water Quality Modelling Using Artificial Neural Network Technique," Aquatic Procedia, vol. 4, pp. 1070-1077, 2015.

[11] A. P. Piotrowski, M. J. Napiorkowski, J. J. Napiorkowski, and M. Osuch, "Comparing various artificial neural network types for water temperature prediction in rivers," Journal of Hydrology, vol. 529, no. 1, pp. 302-315, Oct. 2015.

[12] A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Application of artificial neural networks for water quality prediction," Neural Computing and Applications, vol. 22, no. 1, pp. 187-201, Apr. 2012.

[13] L. Zhang, G. X. Zhang, and R. R. Li, "Water quality analysis and prediction using hybrid time series and neural network models," Journal of Agricultural Science and Technology, vol. 18, no. 4, pp. 975-983, Dec. 2015.

[14] M. J. Alizadeh and M. R. Kavianpour, "Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean," Marine Pollution Bulletin, vol. 98, no. 1, pp. 171-178, 2015.

[15] A. A. Nadiri, S. Shokri, F. T.Tsai, and A. A. Moghaddam, "Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model," Journal of Cleaner Production, vol. 180, no. 7, pp. 539-549, Jan. 2018.

[16] M. Khadr and M. Elshemy, "Data-driven modeling for water quality prediction case study: The drains system associated with Manzala Lake, Egypt," Ain Shams Engineering Journal, pp. 1-9, Sep. 2016.

[17] A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," Journal of King Saud University - Engineering Sciences, vol. 29, no. 3, pp. 237-243, Jul. 2017.

[18] S. Tiwari, R. Babbar, and G. Kaur, "Performance evaluation of two ANFIS models for predicting water quality Index of River Satluj (India)," Advances in Civil Engineering, vol. 2018, pp. 1-10, Mar. 2018.

[19] Y. Khan and S. Chai, "Ensemble of ANN and ANFIS for water quality prediction and analysis - a data driven approach," Journal of Telecommunication, Electronic and Computer Engineering , vol. 9, no. 2, pp. 117-122, 2017.

[20] J.-S. R. Jang, "Frequently Asked Questions - ANFIS in the Fuzzy Logic Toolbox," [Online]. Available: http://www.cs.nthu.edu.tw/~jang/anfisfaq.htm.

[21] A.Khashei-Siuki and M. Sarbazi, "Evaluation of ANFIS, ANN, and geostatistical models to spatial distribution of groundwater quality," Arabian Journal of Geosciences, vol. 8, no. 2, pp. 903-912, Nov. 2013.

[22] N. Q. Sultan, S. Saeed, K. Salar, J. Salar, K. Ozgur, and S. monthly pan evaporation using wavelet support vector regression and wavelet artificial neural Engineering Applications of Computational Fluid Mechanics, vol. 13, no. 1, pp. 177-187, Jan. 2019.
[23] A. A. Najah, A. El-Shafie, O. A. Karim and O. Jaafar, "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation," Neural Computing and Applications, vol. 21, no. 5, pp. 833-841, Nov. 2010.
[24] R.Barzegar, J. Adamowski, and A. A. Moghaddam, "Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran," Stochastic Environmental Research and Risk Assessment, vol. 30, no. 7, pp. 1797 1819, Jan. 2016.
[25] Z. Fu, J. Cheng, M. Yang, and J. Batista, "Prediction of industrial wastewater quality parameters based on wavelet-ANFIS model," in Proc. IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, 2018.