DESIGN AND DEVELOPMENT OF EFFICIENT WATER QUALITY PREDICTION MODELS USING VARIANTS OF RECURRENT NEURAL NETWORKS

Jitha P Nair[1]*, Vijaya M S[2]

Abstract:
Numerous pollutants have exerted a major impact on water quality in recent years, and the health of all living organisms and the environment is directly affected. The most effective water management indicator is Water Quality Index (WQI) developed by BIS (2004). The prediction and modelling of water quality are essential in finding the pollution source and treating it effectively. This study aims to build an efficient river water quality indicator prediction model and classify indicator values according to the Indian Drinking Water Standards (BIS 2004). Data were collected from 11 sampling stations at different points on the Bhavani River in Kerala and Tamil Nadu. The Water Quality Index is computed by using the 28 different parameters that affect the quality of water. The feature selection and data normalisation are applied to develop an efficient river water quality dataset. The WQI prediction model is built using deep learning architectures such as GRU, LSTM, and RNN. The performance of the deep learning based WQI prediction models are compared with traditional machine learning based models. The performance analysis indicates that the GRU-based prediction model shows promising results in predicting water quality.

Keywords: Exploratory Data Analysis, Prediction Model, River Water Quality, Machine Learning algorithms, Deep Learning Architecture.

[1]*Research Scholar, Department of Computer Science PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India Email: jithapnair20@gmail.com
[2]Associate Professor, Department of Computer Science PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India

***Corresponding Author:** Jitha P Nair
*Research Scholar, Department of Computer Science PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India *Email:* jithapnair20@gmail.com

DOI: - 10.31838/ecb/2023.12.si5.0143

1. Introduction

The survival and existence of life on earth require water, and the main source of it are rivers and streams, but they are being polluted day by day as a result of hazardous wastes and pollutants. Water pollution is becoming a dangerous issue for the entire life on earth. Due to the scarcity of fresh water, a large portion of the population today depends on groundwater for drinking, agriculture and industry. Many people rely on groundwater sources such as hand pumps, bore wells and dug wells for water. Groundwater quality is currently declining due to various factors such as host rock composition, rock-water interactions, soil matrix, climate change, and groundwater depth. The use of contaminated water increases the risk of infants developing diarrheal sickness, which raises child mortality rates. According to the World Health Organisation (WHO) (Wang et al., 2017a), an estimated 1.1 billion people worldwide require access to clean drinking water, and 2.6 billion require basic sanitation. The main reason for water contamination is because of the increase in population, automobile, and industry growth. Water purity factors including temperature, turbidity, pH, and other factors are impacted by sewage waste from pollution sources. This increases the spread of illnesses that are transmitted through water and results in the demise of aquatic creatures and plants. In the current situation, pollution disrupts the food chain and wrecks the ecology. People have used a variety of research methods to assess and monitor the water quality, labelling the water quality index. Several researchers spent the majority of their time finding the best model for determining the quality of water.

The prediction of water quality (Wang et al., 2018b) is a primary research topic in water environmental issues, as well as the foundation of water resource management and water pollution prevention and control. Continuous water monitoring can be costly and time-consuming. Water quality monitoring is estimated in various regions through a time-consuming process. By using a mathematical model to anticipate the concentration of key contaminants in river water, it is possible to understand the recent trends in river water quality and provide a standard for managing water quality as well as developing and utilising water resources (Slaughter et al., 2017). Forecasting water quality has promise for effective water management.

The water quality index (WQI), which is based on water quality levels, was recommended by the research community as a global monitoring standard for water quality. WQI ratings vary from 1 to 200; a lower number denotes better water quality. Generally, a WQI level of 30 indicates clean water, a level of 30 to 80 indicates slightly polluted water in a river, and a score of 100 or more indicates that the water is deemed polluted. This first phase of building a WQI prediction model consists of the collection of water samples from proper storage, the site and transportation of these samples to laboratories for testing. Machine learning techniques for water quality forecasting have gained the interest of the research community due to their ability to learn water quality patterns over time.

Artificial neural networks, support vector regression, ARIMA model, deep learning methods such as LSTM, Deep Bi-SSRU Learning Network, recurrent neural network, and many other techniques have also been introduced. According to the published, dissolved oxygen (DO) is important because it provides information about the compound, physical, and biotic properties of water. DO, for example, represents the presence of oxygen (O2) molecules in water in terms of mg/L concentration. DO concentration is an important parameter in predicting water quality.

Several studies on water quality prediction in Indian rivers have recently been published. (Kisi & Parmar, 2016) used SVM and an adaptive regression model to predict water pollution in the Yamuna River. The prediction of water quality is important in the current scenario, to diagnose the actual problem it is essential to predict the future pollution level.

(Dohare et al., 2014) conducted a groundwater study in and around Indore. WQI was determined by collecting samples from all wards in the study area for physicochemical analysis. They concluded that during the wet season, most water quality parameters are slightly higher than during the dry season.

(Kannan et al., 2005) investigated recent groundwater quality in the Thanjavur district, determining the spatial distribution of groundwater quality parameters such as TDS, pH, TH, EC, Cl, and NO3, and generating a groundwater quality region map. The majority of the collected samples were insufficient to meet the WHO and ISI drinking water quality standards.

(Zhang et al., 2014) examined the chemical properties of water samples collected from 39 sampling stations prior to the 2011 summer season irrigation period employing geostatistical methods

and multivariate statistical analysis. Principal Component Analysis (PCA) and two modes of cluster analysis were used to identify the factors that influence the composition of groundwater. PCA was used to discover the factors influencing the evaporation effect and the parameters that influence it.

(Krishan et al., 2016) calculated the Ground Water Quality Index (WQI) using data from 27 samples collected in the Rajkot district of Gujarat and seven parameters such as pH, Chlorides, Total Hardness, Total Dissolved Solids, Fluoride, Nitrate, and Sulphate. The study area's maximum and minimum WQI values were 98 and 27, respectively. According to the calculated WQI, 51.8% of groundwater samples were good and 48.2% of groundwater samples fell into the poor category, indicating that they were unfit for human consumption and would require treatment. After the appropriate treatment, the water can be used for human consumption.

(Ezhilarasi & Senthilkumar, 2018) assessed groundwater quality and carried out the analysis in different wards of Coimbatore City using GIS and WQI. The Water Quality Index assists them in understanding the condition of groundwater in the area.

The focus of this research work is to predict the river water quality index of the Bhavani River which flows through two states such as Kerala and Tamilnadu using the deep learning approaches. The research work aims to predict the water quality index using the daily average values of the river water parameters such as turbidity, pH, COD, temperature, BOD, boron, etc collected from monitoring stations from 2016 to 2020. The WQI prediction is modelled as a regression task in this research work and is investigated by employing GRU, LSTM, and RNN deep learning architectures.

2. Water Quality Standards and the Prediction

Forecasting water quality is essential for preserving the ecosystem. Both point sources and nonpoint sources are used to discharge pollution into the river. A typical strategy for managing point source discharges is to impose regulations that outline the maximum permissible pollutant loads or concentrations in runoffs from point sources like stormwater outfalls, municipal wastewater plants, or industry. The most challenging issue is to control nonpoint sources, as it includes agricultural runoff or atmospheric deposition, making it challenging to apply effluent limits to these pollutants. When compared to loadings from point sources, pollutants from non-point sources are significantly higher.

In order to ensure the water quality criteria are met, an ambient water quality management programme strives to develop acceptable water quality standards in water bodies absorbing pollution loads. The river basin's hydrologic, biological, and land use circumstances, the receiving water source's potential uses, and the ability to establish and sustain water quality standards.

Computation of WQI

The cumulative effect of water quality standards on the overall quality of the water is represented by the Water Quality Index (WQI). Converting complex water quality data into information that is clear and useful is the main objective of the WQI.

The parameters for evaluating water quality must be determined in accordance with a defined standard, such as the Indian Standard for Drinking Water Specification (BIS 2004).
The computation of water quality index is calculated using the following steps.

1. Assign weights to all water quality parameters based on their relative importance. The computation of the relative weight (Wi) of each parameter using the following equation: Wi = K/Si

where Wi is the relative weight, K is the weight of each parameter and Sn is the permissible limit.

2. Assign the quality of water rating (Qi) for each parameter: Qi = (Vi / Si) x 100

where Qi is the quality of water rating, Vi is the mean concentration value for each parameter and Sn is the desirable limit as given in the Indian drinking water standard (BIS 2004).

3. The WQI is calculated by first determining the sub-index (SI) for each water quality parameter: Wi x Qi = SI
   SI is the sub-index of the parameter; Qi is the rating based on the parameter concentration.

4. The summation of the SI of each water quality parameter is the WQI.

The characteristics of the water quality parameter are analysed in accordance with BIS drinking water quality requirements. Table.1 displays the BIS water quality parameters permitted limits as well as the formula used to calculate the water quality

index.

Table 1. Computation of Water Quality Index with Permissible Limits

Parameters	BIS Standard (Si)	1/Si	K= ½∑¹⁄Si	Wi= K/Si	Ideal Value	Mean Value (Vi)	Qi= Vi/Si *100	SI= Wi*Qi
Temp	28	0.03	0.118	0.004	0	28	40	0.169
pH	8.5	0.11	0.118	0.013	7	7.3	85.88	1.202
Conductivity	150	0.006	0.11897	0.00079	0	65	43.33	0.03437
Hardness	100	0.01	0.118979	0.00118	0	9	9	0.010708
Sodium	200	0.005	0.118979	0.00059	0	7	3.5	0.002082
TSS	300	0.0034	0.11897	0.00039	0	300	100	0.03965
BOD	3	0.334	0.118979	0.03965	0	2.3	76.667	3.04059
Nitrate-N	0.503	1.98807	0.11897	0.23654	0	0.902	179.3	42.4173
TC	100	0.01	0.11897	0.00118	0	60	60	0.071387

The impact of each parameter on water quality and health implications must be considered while choosing the water quality parameter and the water quality index range which is given in Table 2.

Table 2: BIS (2004) Water Quality Standards

Water Quality Index Value	Water Quality Index Class	Water Quality Label
>121	E	Unsuitable
91-120	D	Very Poor
61-90	C	Poor
31-60	B	Good
0-30	A	Excellent

The water quality prediction requires sufficient quality data for building the prediction model. The standard parameters determining water quality index such as temperature, conductivity, turbidity, total alkalinity, pH, chloride, phenolph thalein alkalinity, ammonia, total Kjeldahl nitrogen, chemical oxygen demand, hardness, mg. hardness, ca. hardness, sulphate, sodium, ca. hardness, total dissolved solids, phosphate, total suspended solids, fixed dissolved solids, boron, potassium, biological oxygen demand and predicted dissolved oxygen are used in this research. They have been collected from the monitoring stations to prepare the river water quality dataset.

3. Data Acquisition and Preparation

Water quality assessment is an effective way to identify and address water contamination. Water quality is assessed and to determine the water quality index depends on the physical, chemical and biological parameters of water. Hence, in this study, the data to predict the water quality are collected from sampling stations of river Bhavani.

3.1 Data Collection

The Bhavani River flows through Tamilnadu and Kerala, India. The river originates from Nilgiri hills, then enters silent valley national park, Kerala and flows through Tamilnadu. The data are collected from eleven sampling stations of the Bhavani River which include Thavalam, Kottathara, Elachi Vazhi, Chalayur, Cheerakuzhy, Karathur, Badrakaliamman kovil, Bhavanisagar, Sirumugai, Bhavani, Sathyamangalam.

The data collected from eleven stations include temperature, conductivity, turbidity, total alkalinity pH, chloride phenolphthalein alkalinity, ammonia, total Kjeldahl nitrogen, chemical oxygen demand, hardness, mg. hardness, sulphate, sodium, ca. hardness, total dissolved solids, phosphate, total suspended solids, fixed dissolved solids, boron, potassium, biological oxygen demand, predicted dissolved oxygen, longitude and latitude to predict the water quality. Fig1. The flow of River Bhavani is depicted below, the area covered for the research.

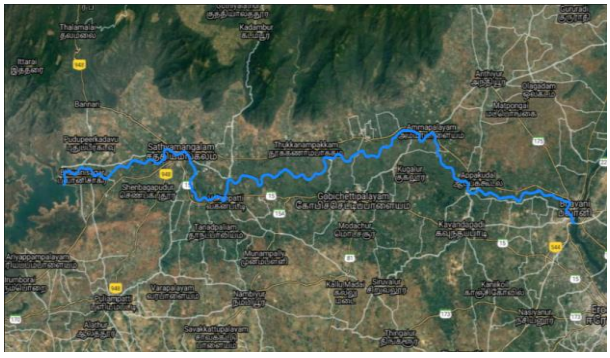


Fig.1. The Flow of River Bhavani

The amount of oxygen dissolved in water may affect the aquatic living creatures, if the level of oxygen is very lower than the metabolic rates of organisms, illnesses, timing of their reproduction, and even migration affects badly. The number of suspended particles in the water is measured by turbidity. E. coli, total coliform bacteria, and faecal coliform bacteria are all to be signs of faecal matter contained in water. The water quality parameters collected from sampling stations are mainly categorised into physical, chemical and biological parameters as shown in Table 3a, Table 3b and Table 3c.

Physical Water Quality Parameters

The physical parameters include temperature, conductivity, turbidity, total suspended solids, fixed dissolved solids and total dissolved oxygen. The river water temperature is heavily influenced by future changes in air temperature and other meteorological and physical factors. The capacity of water to transmit or conduct an electrical current is determined by its electrical conductivity (EC). Turbidity, a term used to describe the cloudiness of water and a metric for how easily light can pass through it, is caused by particles suspended in the water, including silt, clay, organic matter, plankton, and other particles. The TSS and TDS residue that remains after being heated to dryness for a set period of time and at a predetermined temperature is referred to as fixed solids. Some of the physical parameters used in the research work are shown in Table 3a.

Table 3a. Physical Water Quality Parameters

Sl.no	Parameters	BIS Standard (Sn)
1	Temperature	28
2	Turbidity	5
3	Conductivity	150
4	TSS	300
5	TDS	1000
6	FDS	200

Chemical Water Quality Parameters

The chemical parameters of water quality include pH, ammonia, alkalinity, chloride, potassium, sulphate, nitrogen, fluoride, hardness, dissolved oxygen, biological oxygen demand, and chemical oxygen demand. A pH scale of 0 to 14 is used, with 7 representing neutrality. A solution is considered acidic if its pH is below 7, and if its pH is above 7 it's a base. The ammonia concentration in river water is affected by the dead and decay of plants and animals, algal growth, and faecal matter and the increase in the level of ammonia increases water pollution. The alkalinity of water is assessed in order to calculate the quantity of lime and soda needed for water softening. It is the sum of all soluble solids based on acid-neutralizing capacity. Groundwater, lakes and streams, naturally contain chloride the occurrence of relatively high chloride concentrations in freshwater is a sign of water contamination. Numerous sources, such as agricultural runoff, wastewater, and chloride-containing rock, can also provide chlorides to surface water. Magnesium or sodium sulphate deposits found in nature frequently leach, resulting in high sulphate concentrations in natural water. When nitrate levels in surface water are too high, algae can grow quickly and degrade the quality of water. Chemical fertilisers may discharge nitrates into groundwater when used in farming activities. The characteristics of heavily mineralized waters are referred to as hardness. Dissolved oxygen (DO), a major indicator of water pollution, is one of the most important components of water quality in rivers, streams, and lakes. The higher the dissolved oxygen concentration, the better the water quality. Some of the chemical parameters used for the work are included in Table 3b.

Table 3b. Chemical Water Quality Parameters

Sl.no	Parameters	BIS Standard (Sn)
1	pH	8.5
2	Ammonia	50
3	Alkalinity	200
4	Chloride	250
5	Potassium	2.5
6	Sulphate	200
7	Nitrate	0.503
8	Fluoride	1.5
9	Hardness	100
10	DO	7.5
11	BOD	3
12	COD	10

Biological Water Quality Parameters

The biological water quality indicators include total coliform and faecal coliform. The presence or absence of living organisms may be one of the most useful indicators of water quality. The human body maintains a normal population of microbes in the intestinal tract, the majority of which are coliform bacteria. The wastewater contains millions of microbes per millilitre, most of which are harmless. The total coliform and faecal coliform are measured as the total number of colony-forming units (CFUs) in 100 mL of water. Biological water quality parameters used in the research are shown in Table 3c.

Table 3c. Biological Water Quality Parameters

Sl.no	Parameters	BIS Standard (Sn)
1	TC	100
2	FC	60

The river water quality dataset includes the following 26 different physicochemical characteristics with 10560 instances from January 1st, 2016 to December 31st, 2020. The water quality index must be calculated for each parameter with allowed values and unit weight. The water quality index value for each sample was calculated using the Indian Standard for Drinking Water Specification (BIS 2004) and assigned to the corresponding occurrence in the dataset. After computing the WQI and giving class labels to each instance to construct labelled tuples, the water quality index class is established. Finally, a river water quality dataset containing 31 attributes including physiochemical parameters (26), longitude, latitude, station ID, date and calculated WQI and 10560 instances has been created to facilitate supervised learning. As shown in Table 4, time-series data are used to characterise the values of each water quality parameter over a period of time.

Table 4: Sample Physiochemical Parameter Data Collected from Monitoring Stations

Date	10/08/2016	25/05/2017	12/10/2018	16/12/2018	23/12/2019	11/12/2019	08/03/2018	15/06/2018	11/08/2018	18/11/2018`	09/02/2018
Temp	27	26	24	26	22	26	25	27	29	23	22
pH	7.05	7.34	6.08	7.07	7.12	8.27	7.06	7.82	7.47	7.13	7.74
Cl	21	21	21	11	13	9	14	33	22	52	19
COD	4	3.9	4	4	4	4	8	8	16	24	11
TSS	290	280	310	78	67	70	5	10	28	14	11
Fluoride	0.12	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.18	0.18	0.18
Nitrate-N	1.10	1.10	1.10	1.00	1.20	1.00	1.20	1.20	1.20	1.20	1.10
Sodium	27.10	27.10	27.20	27.20	27.00	27.10	27.10	27.00	27.10	27.10	27.10
TC	88	79.41	160	147	163	295	130	192	155	1700	191
FC	80	80	80	39	51	18	27	109	56	430	150

3.2 Exploratory Data Analysis and Data Pre-processing

The Exploratory Data Analysis (EDA) is applied to the river water quality dataset to understand the characteristics of the data and to analyse each parameter for its importance in determining the water quality index. Various statistical approaches such as heatmap, boxplot analysis, pair plot analysis and histogram have been used to analyse and understand the distribution of parameter values. Boxplot analysis shows that conductivity and total coliform parameters have a wide range of values. Parameter such as conductivity has a range from 1 to 1200 and total coliform has a range of 10 to 2500. Hence the parameter values are standardised to fall within the nominal range of each parameter. The min-max normalisation is applied to conductivity and total coliform.

The correlation between the parameters such as positive correlation and negative correlation is analysed by using Pearson correlation and is visualised using a heatmap. The parameters such as pH, turbidity, FDS, TSS, boron and TC are negatively correlated with WQI, which is depicted in Fig.2. The box plot analysis of chlorine, conductivity, dissolved oxygen and temperature are illustrated in Fig.3.

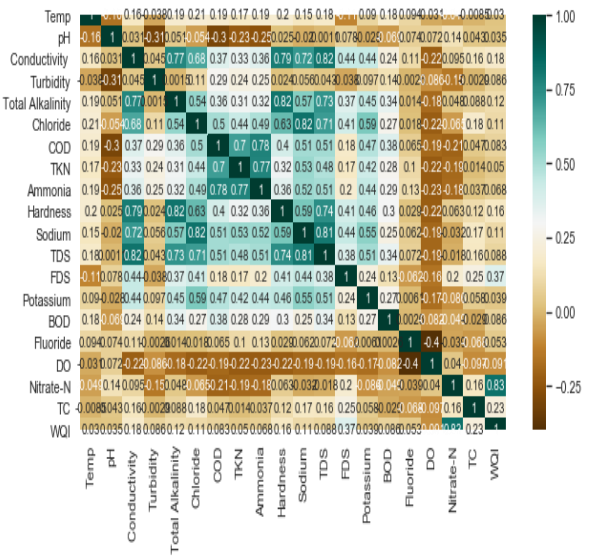


Fig 2. Heatmap Analysis Using River Water Data

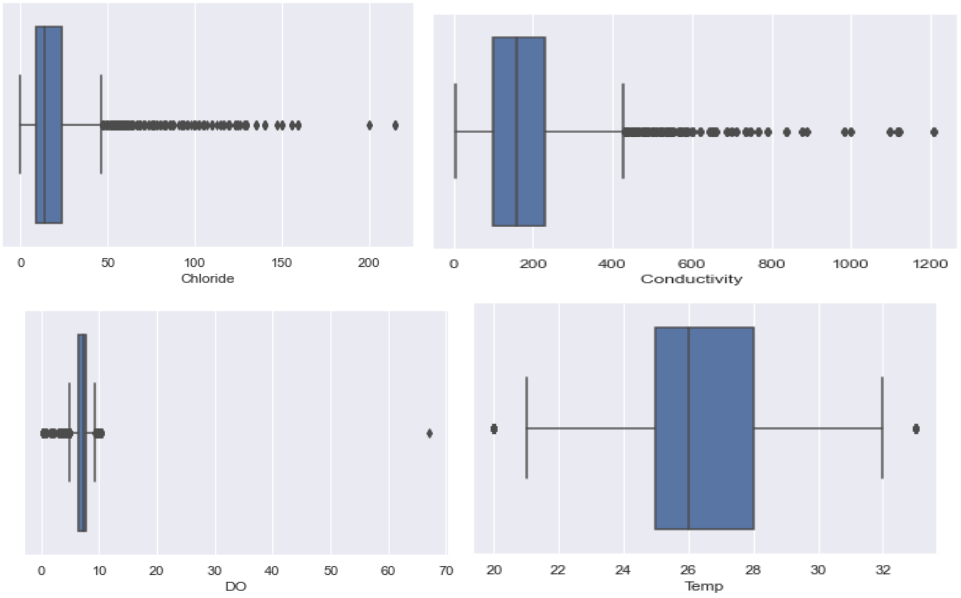


Fig 3. Box-plot Analysis of River Water Quality Data

The analysis results obtained from EDA show that the dataset contains 12 instances with missing values that need to be removed, so data cleaning is carried out. Thus, EDA helps to understand the attribute distributions and correlations between parameters which provide viable solutions for data preparation and data modelling requirements.

3.3. Feature Selection

Feature selection is a crucial phase in predictive modelling in which the appropriate parameters which highly contribute to predicting the target variable are selected. Here the select K best algorithm is used to identify the features which are important in determining the water quality index. According to the select K best feature selection algorithm conductivity has the highest rank in estimating the water quality index whereas phenolphthalein alkalinity, ammonia, and phosphate are in the last ranks. Conductivity, phosphate, fluoride, chloride, alkalinity, sulphate, hardness, sodium, BOD, potassium, DO, nitrate, and coliform, are all important in determining the water quality index. From the developed time series dataset, three attributes were removed as they have no impact on calculating WQI. This feature selection method has resulted in improving the river water quality dataset which finally consists of 10560 instances and 28 attributes for building the river water quality prediction model.

4.WQI Prediction Model using Deep Learning

The problem of water quality index prediction is formulated as a regression problem and solved using deep neural network architectures. Deep neural networks use the data inputs, weights and bias to accurately describe, classify and characterise the data. Deep neural networks have numerous layers of interconnected nodes, with two layers that are visible serving as the input and output layers to enhance prediction. The deep learning model consumes the pre-processed data at the input layer, and at the output layer, the final prediction is made. Large amounts of data can be used to train models, and the model gets better as more data is added and also high-quality predictions when compared to humans.

The structure proposed to produce the WQI prediction model is made up of several basic elements, including 1. Collection of data 2. Exploratory data analysis and data pre-processing 3. Constructing the WQI Prediction Model 4. Model analysis. The river water samples were collected from eleven sampling stations and created the river water quality dataset. The water quality index was calculated and added to the samples to model supervised learning. The exploratory data analysis performed on the river water dataset suggested a few data pre-processing requirements such as normalisation and data cleaning.

WQI prediction models are constructed using deep neural network architectures like RNN, LSTM, and GRU. Various metrics, such as R2 Score, mean absolute error, root mean squared error, and mean squared error has been used to evaluate the performance of the prediction models. Fig.4 depicts the architecture of the proposed WQI prediction model.

Recurrent Neural Network Architecture

Recurrent neural networks are powerful and robust

neural networks which identify patterns in data and make predictions based on those patterns. RNNs operate on the tenet that the output of one layer is saved and fed back into the input to anticipate the behaviour of the next layer. RNN is perfect for machine learning problems requiring sequential input since it has internal memory and uses it to recall its output. RNNs are designed to work with sequential data and use the previous information in the sequence to produce the current output. Each node in RNN layers has the same weights and biases. RNNs have issues with short-term memory and as a result, it causes a vanishing gradient problem. The fundamental challenge for RNN is maintaining data consistency across a large number of time steps, and in a vanilla RNN, the hidden state is constantly being updated. The vanishing gradient problem can be overcome using GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory).

Long Short-Term Memory Architecture

A type of recurrent neural network that has extended memory and prevents disappearing gradients is known as a long short-term memory network (LSTM). A chain of repeating neural network modules in recurrent neural networks with a very simple structure whereas LSTM it has a different structure with four neural networks which interact in a special way. The cell state is the key component of LSTM, which flows directly down to the entire chain with minimal linear interactions. Information can easily be transferred along the cell state without any modification. The LSTM operates entirely on a voluntary basis and can change the cell state by adding or withdrawing information, which is regulated by gates. The LSTM model with large parameters then the model requires more memory and more time to train the model. The LSTM models are very sensitive to different random weight initializations.

Gated Recurrent Unit Architecture

A Gated recurrent unit is an advancement of the standard recurrent neural network and is similar to LSTM. Both LSTM and GRU use gates to regulate the information flow. By storing prior inputs in the internal state of networks and creating target vectors from the history of previous inputs, GRU can process memories of sequential data. GRU replaces the cell state with a hidden state to transfer data. Another difference between GRU and LSTM is that GRU only has two gates: reset and update. The GRU uses update gates and reset gates to solve the RNN vanishing gradient problem. The update gates help the model to determine how far into the future historical data from previous time steps are carried forward. Reset gates are used to determine how much past information is forgotten.

The WQI prediction models can be built using deep learning architectures such as gated recurrent units, long short-term memory and recurrent neural networks, the model can be evaluated using the R2 score, root mean squared error, mean squared error, and mean absolute error.

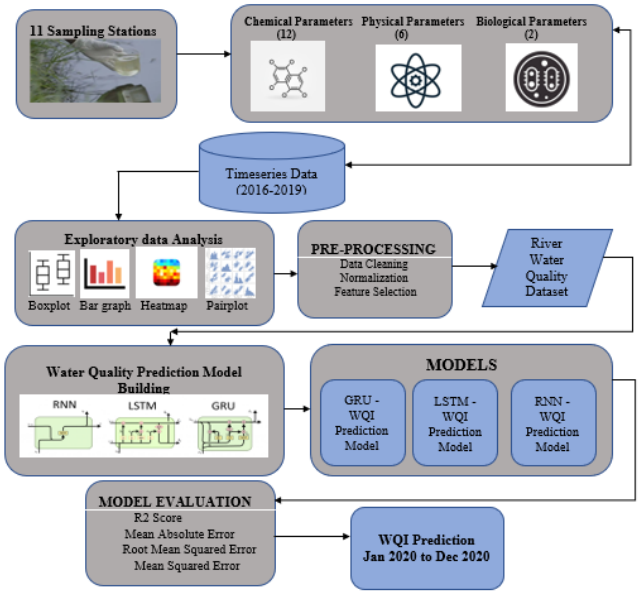


Fig 4. The Framework of the Water Quality Prediction Model

Constructing the WQI Prediction Model

Utilizing the dataset for river water quality, the WQI prediction model is developed using several deep learning algorithms, including GRU, LSTM,

and RNN. Data gathered from various Bhavani River monitoring stations are used to construct the dataset. To comprehend each parameter and how it contributes to the calculation of WQI, exploratory data analysis was used for the dataset and constructed the river water quality dataset. Data normalisation was carried out to parameters such as conductivity and total coliform to standardise the value and also feature selection using SelectK best algorithm to select the important parameters in calculating WQI and removed the less ranked parameters. After completing the pre-processing, the dataset was prepared and divided into 20% of data for testing and 80% of data for training. The model training involves selecting the optimal hyperparameters to improve the efficiency of the model in mapping the input features as independent variables to the target variable as the dependent variable.

Hyperparameters are variables used in model building to improve the accuracy and to fine-tune the WQI prediction model, the hyperparameters used in deep learning architectures are hidden layers, dense layers, optimizer, epoch, momentum, batch size, activation function and dropout. The layers that exist between the input and output layers are known as hidden layers. A dense layer is a layer in which each layer receives input from all layers in the previous and thus, it is densely connected. Dense layers improve overall accuracy and the range is set to 5 to 10 units. Optimizers are techniques used to modify the neural network's properties, such as its weights and learning rate, in order to minimise losses and address optimization issues. The epoch size determines how many complete iterations of the dataset must be run. Momentum is a special hyperparameter that allows the search direction to be determined by the accumulation of gradients from prior steps rather than just the gradient from the current step. Activation functions are used to introduce nonlinearity into the model. This allows deep learning models to learn nonlinear prediction bounds. The activation function can split them into different layers and get a reduced output of the density layer. The dropout layer improves in avoiding overfitting in training by bypassing randomly selected layers, limiting sensitivity to particular layer weights. The learning rate determines the speed at which a deep model replaces an already learned concept with a new one. The WQI prediction model built using deep neural architectures such as GRU, LSTM and RNN have been used and hyperparameters are employed to improve the efficiency of the model and the performance of the models is evaluated.

Model Analysis

To determine the optimal model, the effectiveness of the proposed models in forecasting the water quality index is assessed. The performance of the WQI prediction models is evaluated using the metrics such as R2 score, root mean squared error, mean squared error, and mean absolute error.

The R2 score value determines the accuracy of the model. If the R2 score value is high then the model is predicting the target variable efficiently and if the R2 score is less than 0.5 then the model is not predicting accurately. The R2 score can determine how much better a regression line is than a mean line. $R2score = 1 - \frac{ssr}{ssn}$, where ssr is the squared sum error of the regression line and ssn is the squared sum error of the mean line.

The mean absolute error is a straightforward statistic for calculating the absolute difference between actual and anticipated values. $MAE = \frac{1}{N}\sum abs(Ya - Yb)$, where Ya is the actual output value and Yb is the predicted output values.

The square root of the mean squared error is the root mean squared error. If the RMSE number is large, the model is not an efficient prediction model; a lower error rate indicates that the model is an efficient model for predicting the target variable.$RMSE = \sqrt{1}/n\sum(Ya - Yb)$,

where Ya is the actual output value, Yb is the predicted output value and n is the number of data points.

The mean squared error is a popular and concise metric with a slight difference from the mean absolute error. The squared difference between the actual and predicted values is calculated using mean squared error. The actual output value is Ya and the predicted output value is Yb as depicted, $MSE = 1/n\sum(Ya - Yb)^2$, where n is the number of data points.

The prediction models are found to be effective when the error rate is less with a high R2 score value. In this work, the performance of the WQI predictive models is evaluated using the above metrics with 20% of the dataset as the test set.

5.Experiment and Results

The experiments have been carried out by training the Bhavani River water dataset using deep learning algorithms such as GRU, LSTM, and

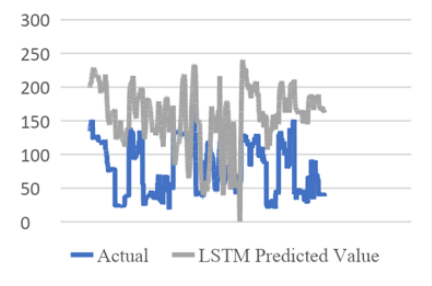
RNN, and implemented using python libraries. The training dataset with 80% of the instances of the river water dataset covering 8124 tagged samples has been used for training the networks. The test set with 2009 instances has been used for testing the performance of the prediction models and evaluated for its efficiency in forecasting the water quality using the R2 score, root mean squared error, mean squared error, and mean absolute error.

The deep neural architectures GRU, LSTM, and RNN are defined with various hyperparameters such as dense layer values from 5 to 10 units, optimizer as Adam optimizer. The epoch size was given as 20, 50,100,150 and 200 epoch size. The momentum is set from 0.5 to 0.9, the activation functions are defined with both on and off. The batch size is fixed as either 32 or 64, the dropout unit is 0.2 and the learning rate is 0.1. Various hyperparameters used to fine-tune the WQI prediction model are shown in Table 5.

Table 5. Hyperparameters Setting

Hyperparameter	Values
Dense Layer	5to 10 units
Optimizer	Adam
Epoch	20,50,100,150,200
Momentum	0.5 to 0.9
Activation Function	On/Off
Batch size	32, 64
Dropout	0.2
Learning rate	0.1

The experimental results with respect to the deviation between the predicted values and actual values shown by GRU, LSTM, and RNN WQI prediction models are illustrated in Fig. 5a,5b and 5c.
From the figures, it is found that the deviation between the actual values and the predicted values in the case of the GRU prediction model is less with the threshold value when compared with LSTM and RNN.



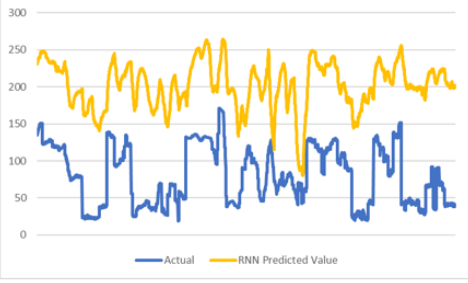


Fig.5a. WQI LSTM Prediction vs Actual Value **Fig.5b.** WQI RNN Prediction vs Actual Value

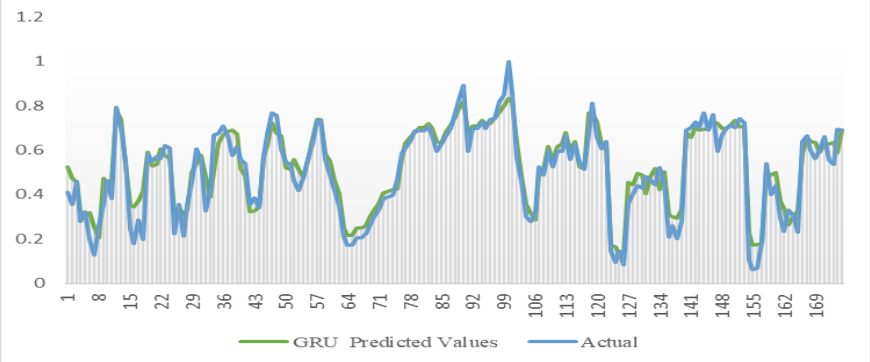


Fig 5c. WQI GRU Prediction vs Actual Value

The R2 score value of GRU based WQI prediction model shows 0.885 and is high when compared to other prediction models. The R2 score value of the RNN prediction model yields 0.828 and the LSTM prediction model is 0.852 with an epoch size set to 200 which is illustrated in Fig.6a.

The WQI prediction results showed the least mean absolute error value of 0.364 for the GRU prediction model with epoch size 200, 0.483 for the LSTM prediction model and 0.496 for the RNN prediction model.

The regression model results of prediction results observed that the root mean squared error value of the GRU prediction model trained with epoch size 200 is 0.348, the LSTM prediction model is 0.5832 and the RNN prediction model is 0.525.

The comparative performance results of the deep learning model concerning the metrics mean absolute error, root mean squared error and R2 score variation is illustrated in Fig. 6.

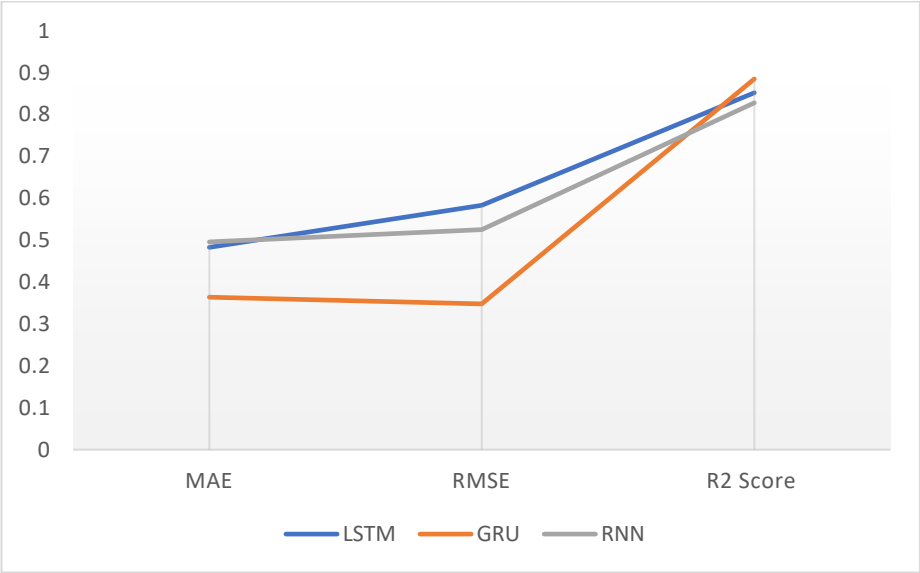


Fig. 6. Performance analysis of Deep Learning Prediction Models

The performance of the deep models is also compared with traditional machine learning algorithms like random forest, linear regression, support vector regressor, and MLP regressor.

Regression model prediction results show that the GRU prediction model trained with the least mean absolute error value with epoch size 200 is 0.364 and for MLP regressor obtains a 0.714 error value. The WQI prediction results showed that the least root means squared error value of 0.348 for the GRU prediction model with epoch size 200 and the MLP regressor acquires a 0.763 error rate.

The WQI prediction models are built using the Bhavani River water dataset, comparing the R2 score value of deep learning prediction models with traditional machine learning approaches. It is found that GRU based WQI prediction model shows the R2 score value as 0.885 and the MLP regressor obtains 0.7342.

The results of the prediction models evaluated using different metrics such as mean absolute error, root mean squared error, mean squared error and R2 score are illustrated in Fig.7a, Fig. 7b and Fig.7c respectively.

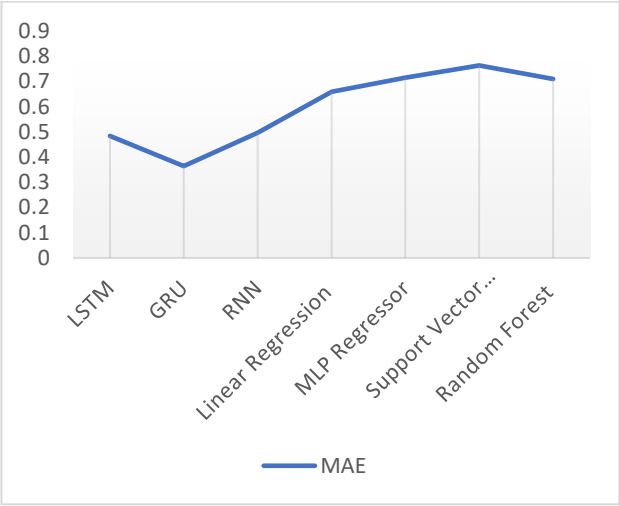


Fig.7a. MAE of Prediction Models

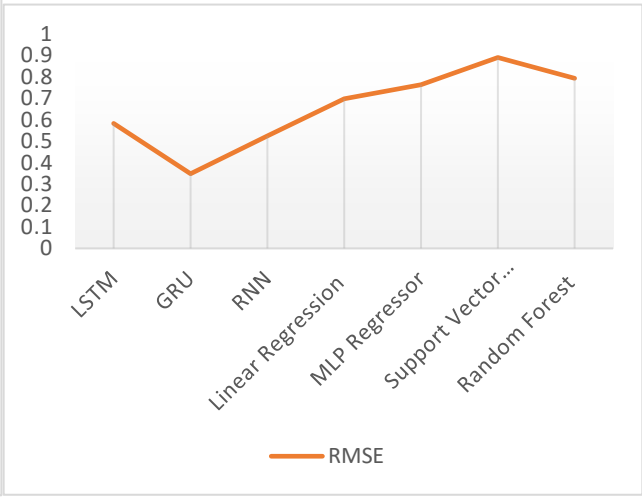


Fig. 7b. RMSE for prediction models

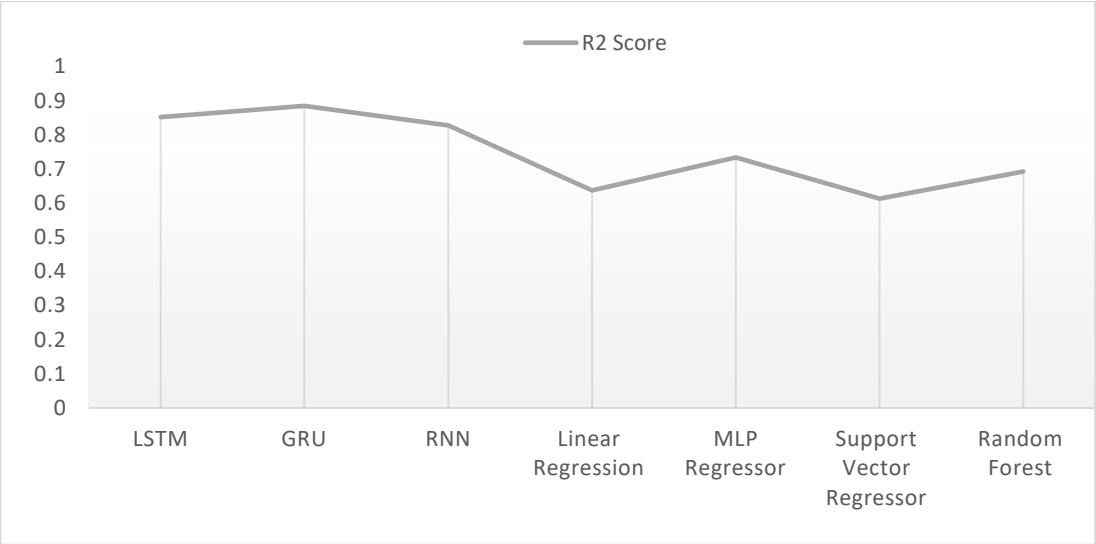


Fig. 7c. R2 score value of Prediction Models

The GRU-based WQI prediction model yields less error rate as compared to all other prediction models used in predicting water quality index using the river water quality dataset. It is proven from the evaluation results that the GRU prediction model yields high accuracy and less error rate. The comparative performance results of the WQI prediction model are shown in Table 6 and the comparative performance analysis is illustrated in Fig 8.

Table 6: Comparative Performance Results of Water Quality Index Prediction Models

Models	MAE	RMSE	R2 Score
LSTM	0.483	0.5832	0.852
GRU	0.364	0.348	0.885
RNN	0.496	0.525	0.828
Linear Regression	0.659	0.698	0.6375
MLP Regressor	0.714	0.763	0.7342
Support Vector Regressor	0.763	0.89	0.6132
Random Forest	0.709	0.793	0.6923

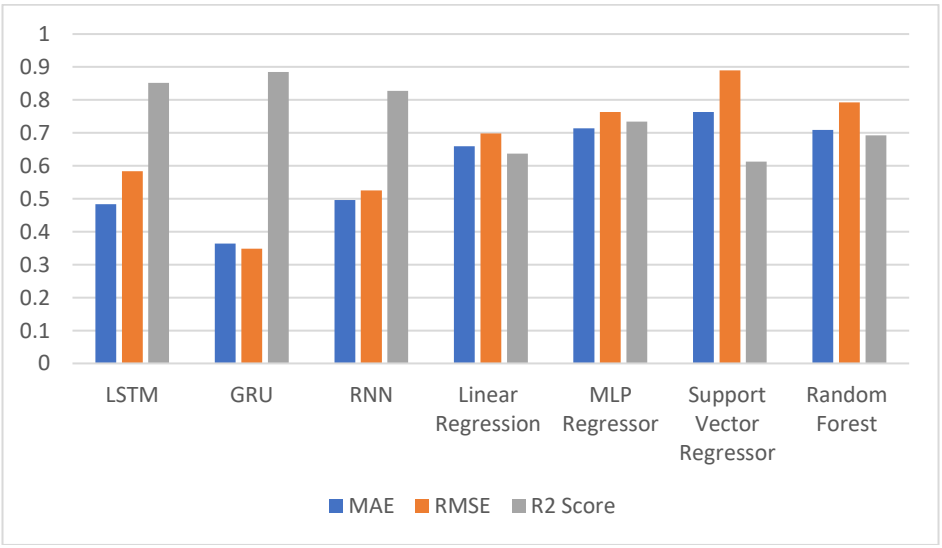


Fig. 8. Performance comparison of Water Quality Index Regression Models

From the comparative performance analysis of various WQI predictive models, it is observed that deep learning-based WQI prediction models show better performance than traditional machine learning algorithms. The machine learning approach is good for building any predictive models like water quality index prediction, but the recent deep learning approach improves the accuracy of the prediction. More powerful deep neural network architectures such as GRU, LSTM

and RNN enhance the recognition of the correlation between target variables and the set of predictors through representation learning. The training of self-learnt features in GRU, LSTM and RNN increases the prediction rate of models. The proper setting of hyperparameters for training the network reduces the error rate of trained models.

The RNN architecture has a gradient vanishing problem in optimising the training, due to which the error rate shown by the model is higher. The GRU models use fewer parameters while training and need only less memory to execute, so it is faster than LSTM and RNN. The GRU-based WQI prediction models perform efficiently and are more suitable for time series-based water quality datasets.

6. Conclusion

This paper demonstrates the implementation of water quality prediction using deep neural network architectures such as GRU, LSTM and RNN. Eleven sampling stations along the Bhavani River yielded river water quality data, including physicochemical characteristics, latitude and longitude were employed in building the WQI prediction model. EDA was applied to the river water dataset to understand the distribution of data and the importance of each water quality parameter in predicting WQI. The deep learning-based WQI prediction models have been developed using GRU, LSTM, and RNN architectures. The GRU prediction model yields high accuracy with less error rate as compared to other algorithms in predicting WQI. In the future, the model efficiency can be improved by adding the seasonal data with the existing physio-chemical properties and fine-tuning the deep learning architecture to predict WQI.

References

1. WangY, Zhou J, Chen K, Wang Y, Liu L. 2017. Water quality prediction method based on LSTM neural network. In 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE. 1-5.
2. Li L, Jiang P, Xu H, Lin G, Guo D, Wu H. 2019. Water Quality Prediction Based On Recurrent Neural Network and Improved Evidence Theory: A Case Study Of Qiantang River, China. Environ Sci Pollut Res 26(19): 19879–19896
3. Liu J, Yu C, Hu Z, Zhao Y, Bai Y, Xie M, Luo J. 2020. Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network. IEEE Access. 8:24784–24798
4. Yahya A, Saeed A, Ahmed AN, Binti Othman F, Ibrahim RK, Afan HA, Elshafie A. 2019. Water Quality Prediction Model Based Support Vector Machine Model For Ungauged River Catchment Under Dual Scenarios. Water.11(6):1231
5. Kisi O, Parmar KS. 2016. Application Of Least Square Support Vector Machine And Multivariate Adaptive Regression Spline Models In Long-Term Prediction Of River Water Pollution. J Hydrol. 534:104–112
6. Wang J, Xiang F, Qiu F, Wang H, Liu H. 2018. Research Progress Of Water Quality Prediction Model. Guild of Environmental Sciences. 37(4):63-67.
7. Slaughter AR, Hughes DA, Retief DCH, Mantel SK. 2017. A Management-Oriented Water Quality Model For Data-Scarce catchments. Environmental Modelling and Software. 97:93-111.
8. Nair JP, Vijaya MS. 2021. Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). 1747-1753.
9. Dohare D, Deshpande S, Kotiya A. 2014. Analysis of Groundwater quality parameters: A Review. Research Journal of Engineering Sciences. 3(5):26-31.
10. Kannan V, Ramesh R, Sasikumar C. 2005. Study on groundwater characteristics and the effects of discharged effluents from textile units at Karur district. Journal of Environmental Biology, 26(2):269-272.
11. Zhang X, Qian H, Chen J, Qiao L. 2014. Assessment of Groundwater Chemistry and Status in a Heavily Used Semi-Arid Region with Multivariate Statistical Analysis. Water. 6(8):2212-2232.
12. Krishan G, Singh S, Kumar CP, Gurjar S, Ghosh NC. 2016. Assessment of Water Quality Index (WQI) of Groundwater in Rajkot District, Gujarat, India. Journal of Earth Science & Climatic Change. 7(3):1000341.
13. Ezhilarasi M, Senthilkumar V. 2018. Geo-Chemical Analysis for Groundwater Quality Using Geospatial Application, International Research Journal of Engineering and Technology (IRJET). 5(4).
14. Jitha P Nair, Vijaya MS. 2022. River Water Quality Prediction and index classification using Machine Learning. Journal of Physics: Conference Series. IOP Publishing. 2325(1).

15. Nair Jitha P, Vijaya MS. 2023. Exploratory Data Analysis of Bhavani River Water Quality Index Data. In: Kumar S, Hiranwal S, Purohit SD, Prasad M. (eds) Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems. Springer, Singapore.
16. Santhana Lakshmi, V., Vijaya, M.S. (2022). A Study on Machine Learning-Based Approaches for PM2.5 Prediction. In: Karrupusamy, P., Balas, V.E., Shi, Y. (eds) Sustainable Communication Networks and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 93. Springer, Singapore.
17. Nair, Jitha P., and M S Vijaya. 'Analysing And Modelling Dissolved Oxygen Concentration Using Deep Learning Architectures'. International Journal of Mechanical Engineering, vol. 7, pp. 12–22.