Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7722400

ENHANCED WATER QUALITY PREDICTION MODEL WITH SEASONAL TIME SERIES DATA

Jitha P Nair[1] and Vijaya M S[2]

[1] Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India

[2] Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Peelamedu, Coimbatore, India

ABSTRACT

In recent years, several contaminants have posed a threat to rivers, streams and lakes. The ability to analyse and anticipate water quality has emerged as an aid in the fight against the contamination of water. Various seasonal factors along with physicochemical properties influence water quality over time. Water quality data becomes time series data and the values of parameters change as meteorological conditions change over seasons at each location. Hence good time series analysis is required to forecast water quality. Considering the significance of Recurrent Neural Network (RNN) for time sequence data, this work is intended to build a water quality prediction model by learning seasonal patterns in the time series dataset. The dataset contains 10560 unique instances that describe both physicochemical and seasonal factors. Predictive models are developed using RNN and its variants GRU and LSTM and evaluated. Promising results are produced as a result of augmenting seasonal data with regular physicochemical properties while training the model.

Keywords: River Water Quality, Prediction Model, Deep Learning Architectures, Physicochemical Parameters, Seasonal Parameters

1. INTRODUCTION

The survival of the vast majority of living species, including humans, depends on water, making it the essential resource for life. Water with high quality is important for all forms of life. Water-dependent species can only tolerate less amount of pollution before they disappear. When some conditions are not satisfied, the survival of these creatures is endangered.

Automatic water quality monitoring stations need to be built in many important areas and accurate water quality prediction methods are critical for timely monitoring and controlling water pollution. Various water treatment methods such as anaerobic, aerobic, activated sludge methods, and membrane bioreactor treatments are used to treat the wastewater and thereby control the pollution. Physicochemical treatment is employed to separate the colloidal particles in the water.

The physicochemical factors such as conductivity, turbidity, total alkalinity, chloride, ammonia, hardness, sulphate, sodium, phosphate, boron, potassium, BOD, fluoride, nitrate, coliform and dissolved oxygen are some of the common parameters to determine water quality index. Water quality is affected by several seasonal conditions during various seasons. In addition to physicochemical properties, seasonal attributes such as temperature, dew, humidity, precipitation, wind speed, and visibility are also equally important in predicting water quality.

Several research works use only a limited set of physicochemical parameters to build water quality, forecasting models. Increasing the number of physicochemical factors and the inclusion of seasonal variables can improve efficiency in water quality prediction.

The grey relational method, mathematical statistics method, model-based approach, Bayesian approach, Genetic Algorithm, MLP regressor, and support vector regressor are computational methods used by researchers currently in the existing water quality prediction research.

Umair Ahmed et al. [2] established an efficient water quality prediction Framework with supervised machine learning. This framework provided a strategy that employs four information boundaries, namely temperature, turbidity, pH, and solids that have been entirely dispersed. The research was supported by data from PCRWR, which included 663 samples from 12 distinct wellsprings in Pakistan's Rawal Lake. WQI was evaluated utilising a range of AI calculations directed by agents and also calculating relapse and categorization. The eight lapse calculations for WQI and 10 classification calculations for ordering experiments into predetermined WQC computations had been assessed.

Shuangyin Liu et al. proposed a half-and-half methodology of help vector relapse with hereditary calculation advancement for hydroponics water quality prediction [5]. This paper proposes a forecast model based on help vector relapse (SVR) to address the hydroponics water quality expectation issue. When putting together a successful SVR model, the SVR boundaries should be set with caution. This study presents a half-and-half methodology known as genuine worth hereditary calculation uphold vector regression (RGA-SVR), which looks for the best SVR boundaries using genuine esteemed hereditary calculations and then uses the best boundaries to build the SVR models. The methodology is used to forecast the hydroponics water quality data collected from Yixing's oceanic plants in China. The results show that RGA-SVR outperforms the standard SVR and back-engendering (BP) neural organisation models based on the root mean square error (RMSE) and mean outright rate blunder (MAPE). This RGA-SVR model is a viable method of dealing with anticipated hydroponics water quality.

Salisu Yusuf Muhammad et al. [6] introduced a machine learning-based water quality classification model. This article proposes a reasonable grouping model for ranking water quality based on AI calculations. The paper separated and analysed the presentation of various arrangement models and calculations to identify the key factors that contributed to the classification of the water nature of the Kinta River in Perak, Malaysia. Five mathematically precise models were evaluated, compared, and displayed. The Lazy model utilising the K Star calculation was the most accurate grouping model among the five models, with an exactness of 86.67%. In general, wastewater is hazardous to human health, and it is essential to develop logical models to address this problem.

Liao and Zhao [18] focused on dissolved oxygen for water quality prediction and proposed a combined model consisting of fuzzy neural networks (FNN), principal component analysis (PCA), and differential evolution by the BP algorithm (DEBP). PCA contributes to the dimension reduction of the input data vector and differential evolution algorithm.

Wang et al. [17] demonstrated an LSTM (long- and short-term memory) neural network-based deep learning approach. The LSTM NN model was constructed for prediction, followed by the collection of training data from Taihu Lake and the selection of appropriate parameters to improve neural network accuracy. Due to the nonlinear, dynamic, changing, and complex nature of the water parameter quality parameters, predicting WQ is a hard task. Due to these traits, traditional forecasting algorithms suffer from poor accuracy and increased processing complexity.

This study aims to develop an enhanced water quality prediction model by exploiting the importance of Recurrent Neural Network (RNN) for time sequence data. Time series seasonal data is obtained from the visual crossing site based on the location of eleven sampling stations of the Bhavani River. The river water quality predictive model is built using RNN variants such as LSTMs and GRUs, and the models are evaluated for their its efficiency.

2. DATA COLLECTION AND DATASET PREPARATION

In our earlier research, a water quality predictive model was constructed by identifying the trends from physicochemical features in time series river water data. The training dataset included 26 physicochemical parameters such as pH, conductivity, turbidity, phenolpth alkalinity, total alkalinity, chloride, COD, TKN, ammonia, Ca. hardness, Mg. Hardness, sulphate, sodium, TSS, TDS, FDS, phosphate, boron, potassium, BOD, fluoride, Nitrate-N, DO, TC and FC. as shown in Table 1.

Table 1: Sample Physicochemical Parameters Collected from Sampling Stations

pH	7.15	7.46	7.5	7.18	7.45	7.05	7.4	7.38	7.56	7.1
Conductivity	340	339	339	340	340	342	341	339	340	340
Turbidity	2	2	2	2	2	2	2	2	2	2
Phenolpth Alkalinity	0	0	0	0	0	0	0	0	0	0
Total Alkalinity	111	110	112	111	110	110	112	111	112	111
Chloride	21	21	22	21	20	20	20	21	21	21
COD	4	3.9	4	3.9	4	4	4	3.9	3.9	4
TKN	0.1	0.1	0.09	0.1	0.1	0.09	0.1	0.1	0.1	0.11
Ammonia	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

Hardness	118	118	119	119	119	119	118	118	118	117.5
Ca. hardness	74	74	74.5	74.5	74	73.5	73.5	73.5	74	74
Mg. Hardness	44	44	44	43.5	43.5	43.5	44	44	44	44
Sulphate	12	12.5	12	12	12.5	12.5	12	12	12.5	12
Sodium	27.1	27.1	27.2	27.2	27	27.1	27.1	27	27.1	27.1
TSS	300	300	300	300	300	300	300	300	300	300
TDS	190	190	189	189	189	190	189	190	189	188
FDS	174	174	174	174.5	174.5	174	174	174	173.5	173
Phosphate	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
Boron	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Potassium	2.67	2.67	2.66	2.66	2.67	2.67	2.66	2.66	2.66	2.66
BOD	0.89	0.87	0.89	0.88	0.85	0.87	0.82	0.81	0.88	0.82
Fluoride	0.12	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.18	0.18
Nitrate-N	1.1	1.1	1.1	1	1.2	1	1.2	1.2	1.2	1.2
DO	6.99	6.97	6.81	7.19	7.3	7.39	7.06	7.02	6.97	7.39
TC	88	98	118	86	65	105	83	113	65	85
FC	80	80	80	79.5	79.5	79	79.5	80	80	80

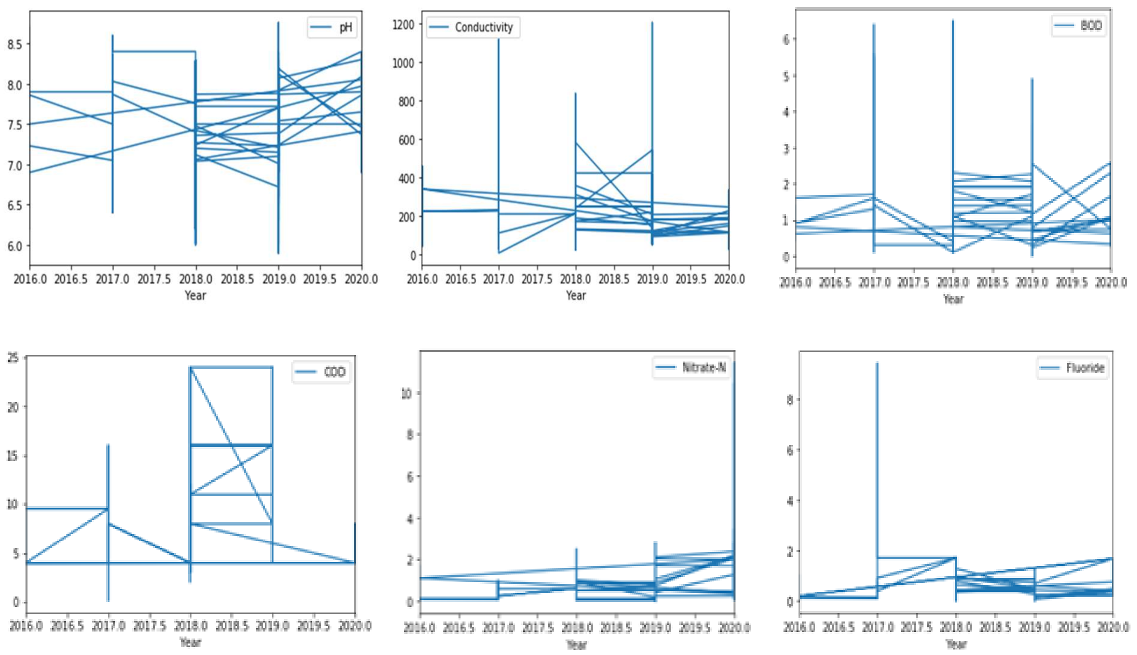
Seasonal parameters affect river water quality over time due to sudden climatic changes. It has been observed from the literature that the seasonal parameters have an impact on the water quality index and its prediction over time series data. Simultaneous rainfall and humidity are strongly related, the relative humidity improves as a result of the evaporation of rainwater. The Davis Cup Anemometer is used to measure wind speed at a height of three metres above the ground and is compared to more conventional measurements taken at a height of 10 metres. An increase in wind speed decreases the transition time between evaporative stages at low-velocity values. Dew is a vital source of river water that greatly impacts the microclimate and the physiological state of the vegetation. Global warming will alter the distribution of precipitation by altering air temperatures and circulation patterns. All of the seasonal factors have an impact on water quality by altering the acceptable limits of physicochemical parameters, hence diminishing water quality. Hence in this work, the seasonal features are considered for the same period of study and their importance is incorporated to enhance the efficiency of the predictive model.

The seasonal features acquired from visual crossing sites are based on sampling station locations from January 2016 to December 2020 and the sample data is shown in Table 2. Seasonal Parameters like dew, humidity, sea level pressure, precipitation, precip over, wind speed, wind direction, cloud cover, and visibility are considered here as these characteristics change with the season over time. These seasonal attributes are pooled with physicochemical parameters to develop a dataset in this work.

Table2: Sample Seasonal Parameters Collected Visual Crossing Site

Temp	25	24	25	25	25	24	24	25	25	25
Dew	15.7	14.6	13.4	13.6	15.6	17.7	18.9	19.4	18.3	17.8
Humidity	59.3	56.72	51.89	53.06	58.8	62.79	68.91	68.63	65.71	63.8
Sea level pressure	1016.6	1017.1	1015.8	1015.7	1014.8	1014.8	1015.5	1015.5	1013.7	1014.5
Precipitation	0	0	0	0	0	0	0.2	0	0	0
Precip cover	0	0	0	0	0	0	4.17	0	0	0
Windspeed	16.3	14.4	13.1	15.4	14	18.7	40.2	13.6	14.4	14.9
Wind dir	52.9	62.3	61.7	68.2	56.5	69.3	114.6	95	94.9	65.1
Cloud cover	27.4	17.9	5.5	14.1	14.6	16	32.3	42.5	26.3	14
Visibility	5.5	6	5.7	5.9	5.6	5.5	4.8	5.3	5.1	5.4

The temporal variation of both physiochemical and seasonal parameters is illustrated in Fig.1a and Fig1 b.



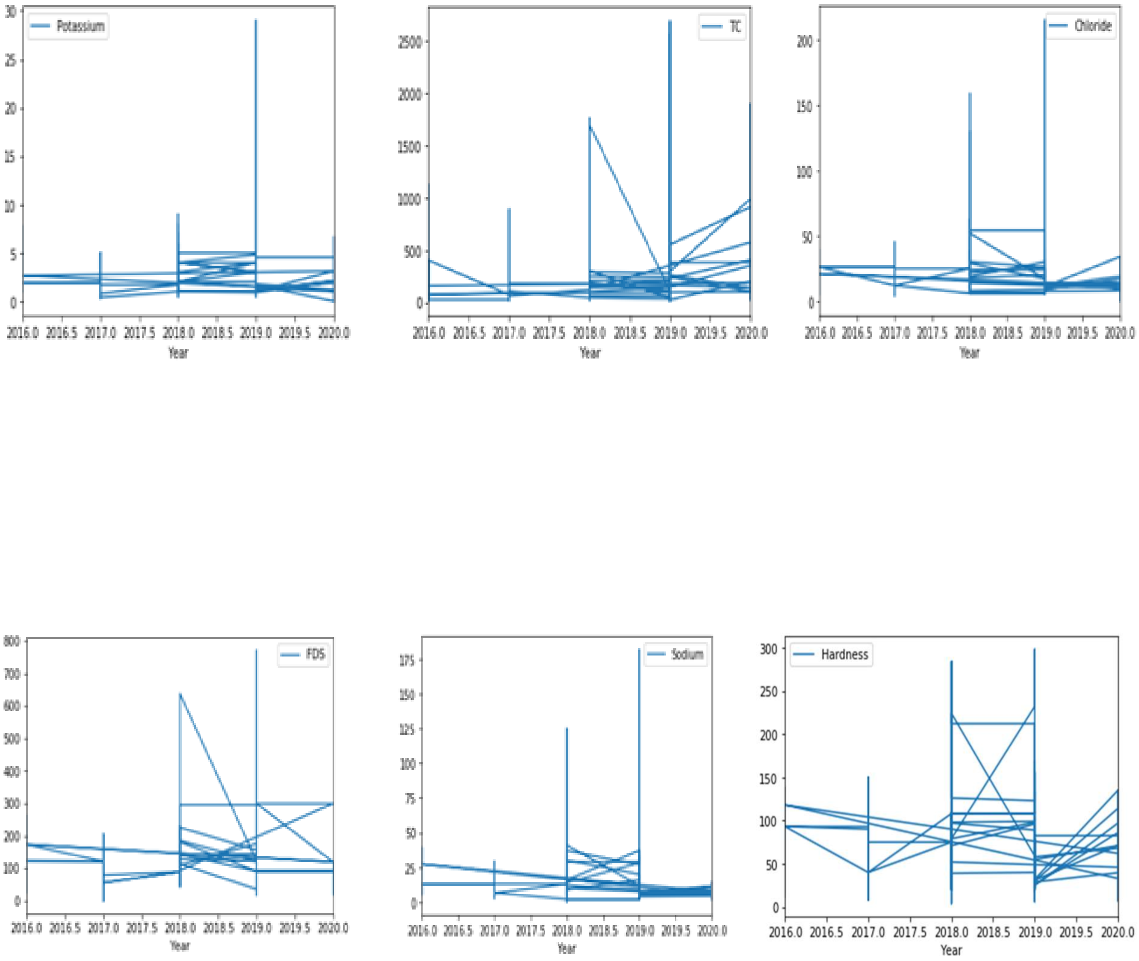


Fig.1 a. Temporal Variation of Some Physicochemical Parameters of Water in the Bhavani River

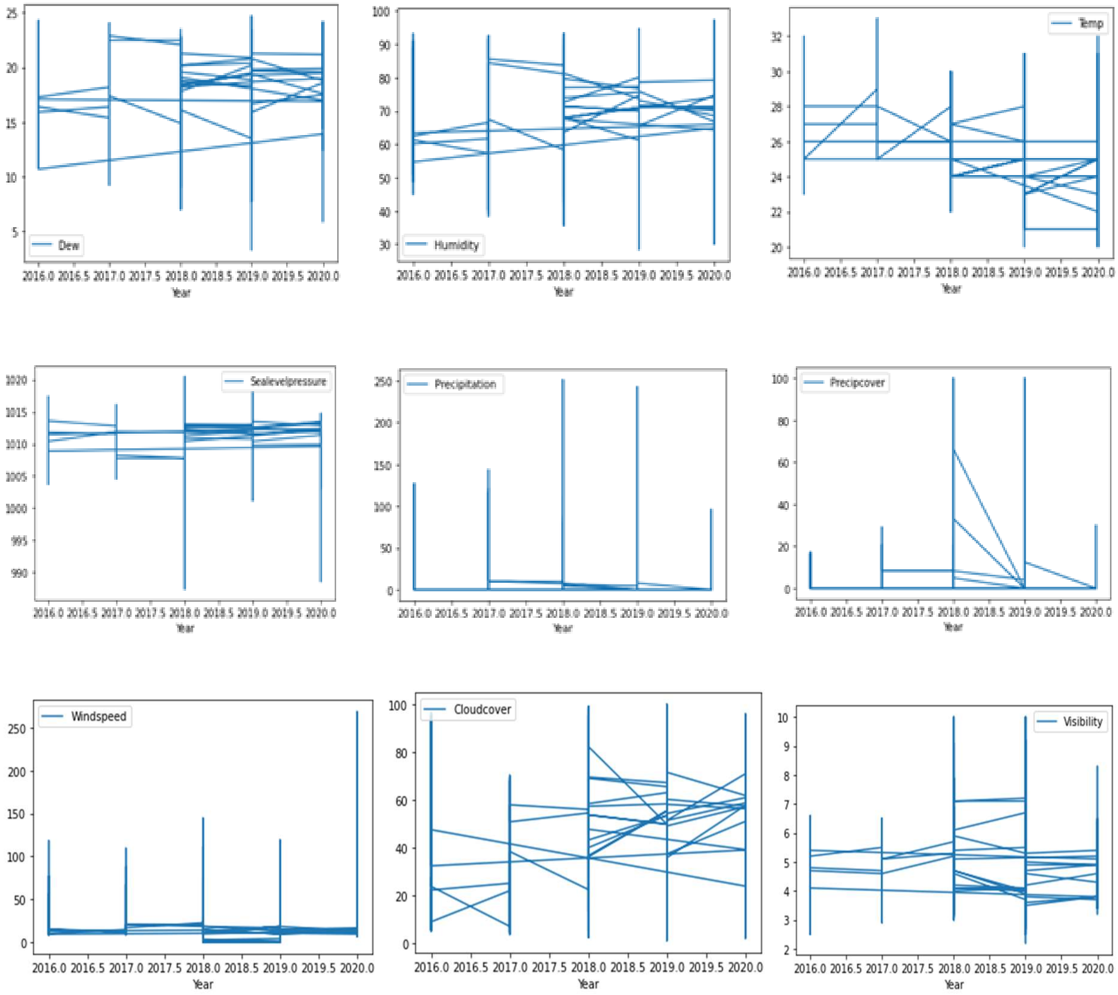


Fig. 1b. Temporal Variation of Some Seasonal Parameters of Water in the Bhavani River

The Water Quality Index (WQI) is a tool used to measure the quality of water. It is composed of several seasonal attributes that are used to assess the overall health of a water body. Temperature is an important attribute as it helps to indicate the presence of certain species, as well as the activity of the water body. Dew is a measure of the amount of water vapour present in the atmosphere. It is significant to determine water quality because it helps to regulate the temperature of the environment, and it can also indicate the amount of precipitation that is likely to occur. Humidity is a measure of the amount of water vapour present in the air. It is also important to measure the water quality as it can affect the rate of evaporation, and also indicate the temperature of the environment.

Sea level pressure is a measure of the atmospheric pressure at sea level and is important in water quality prediction because it can affect the rate at which water evaporates, and also indicate the amount of precipitation that is likely to occur. Precipitation is a measure of the amount of liquid or solid water particles that have fallen from the atmosphere. It is crucial for maintaining water quality as it can affect the amount of dissolved oxygen in the water, and it can also indicate the pollutants that are present. Precip Over is a measure of the amount of precipitation that has fallen over a certain period. It has a significant impact on the quality of water because it can indicate the number of pollutants that are present in the water, and it can also indicate the number of nutrients that are available for plant growth.

Wind speed is a measure of the speed of the wind that is blowing. It is important to water quality prediction because it can affect the rate at which water evaporates, and it can also indicate the pollutants that are present in the water. Wind direction is a measure of the direction in which the wind is blowing. It plays a vital role in water quality because it can affect the rate at which water evaporates, and it can also indicate the number of pollutants that are present in the water. Cloud cover is a measure of the amount of water vapour that is present in the atmosphere. It has a significant effect on water quality due to its ability to affect the rate of evaporation, and it can also indicate the volume of pollutants that are present in the water. Visibility is a measure of how far one can see in the atmosphere. It is a key factor in the quality of water because it can indicate the pollutants that are present in the water, and it can also indicate the number of nutrients that are available for plant growth.

Table3: Water Quality Parameters

Physicochemical Parameters		**Seasonal Parameters**
pH	TSS	Temperature
Conductivity	TDS	Dew
Turbidity	FDS	Humidity
Phenolphthalein Alkalinity	Phosphate	Sea level pressure
Total Alkalinity	Boron	Precipitation
Chloride	Potassium	Precip cover
COD	BOD	Windspeed
TKN	Fluoride	Wind dir
Ammonia	Nitrate-N	Cloud cover

Hardness	TC	Visibility
Ca. hardness	FC	**Spatial Parameters**
Mg. hardness	Dissolved Oxygen	Station ID
Sulphate	**Temporal Parameter**	Latitude
Sodium	Date	Longitude

Thus, twenty-six physiochemical attributes are pooled with ten seasonal attributes along with spatial parameters to develop the WQI-SA dataset. Finally, there is a total of 40 attributes forming the time series data prepared for this research work.

The data collected on river water quality is subjected to exploratory data analysis to comprehend the properties of the data and evaluate the significance of each parameter in generating the water quality index. The physicochemical data collected from the sampling stations and the seasonal data collected from the visual crossing site are listed in Table 3. Statistical methods such as heatmaps, boxplot analysis, pair plot analysis, and histograms have been utilised to analyse and understand the distribution of parameter values. According to boxplot studies, seasonal parameters like wind speed, and cloud cover characteristics have a broad range of values. While wind speed ranges between 10 and 270, and cloud cover is between 0 and 100. Therefore, the parameter values are normalised so that they lie within the usual range for each parameter. Wind speed and cloud cover are standardised using the min-max approach. Using Pearson correlation, the heatmap is used to visualise and analyse the correlation between the parameters, such as positive and negative. The bar graph analysis of humidity, wind speed, cloud cover, visibility and physicochemical parameters are depicted in Figure 2a. pH, turbidity, FDS, TSS, boron, TC, cloud cover, and wind speed are the parameters that have a negative correlation with WQI and are displayed in Fig. 2b.

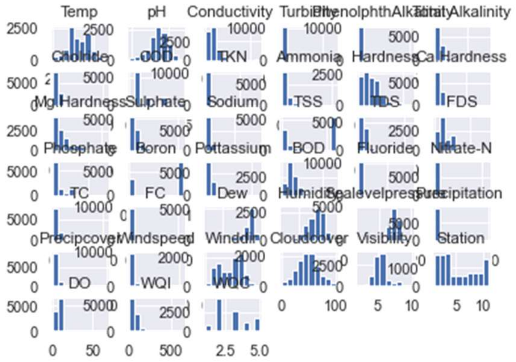


Fig.2a Bar graph analysis

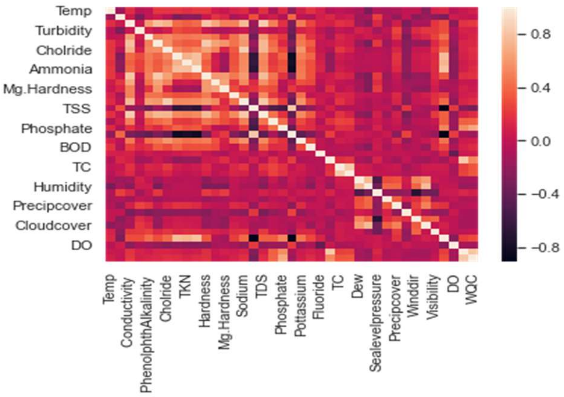


Fig.2b Heatmap analysis

EDA reveals that some instances in the dataset include missing values that must be eliminated, so data cleaning is performed. EDA explains better about the attribute distributions and parameter correlations, providing suitable solutions for data modelling and pre-processing needs.

The Water Quality Index (WQI) is a measure of the overall water quality of the proposed system. The WQI can be used to monitor changes in water quality over time and to assess the suitability of the water body. It is calculated by taking the average of several factors that are indicators of water quality, like dissolved oxygen, pH, nutrient levels, and turbidity. The WQI is then assigned a score based on a range of 0 to 120, with higher scores indicating poor water quality. The WQI is computed and then added as the target variable along with the 40 independent variables for the WQI modelling prediction task. Hence in the work, the dataset includes both physiochemical and seasonal parameters and it contains 10560 instances.

Feature selection is a vital phase in predictive modelling in which appropriate parameters that contribute significantly to predicting the target variable are chosen. The select K best algorithm is employed in this case to identify important features in calculating the water quality index. According to the select K best feature selection algorithm, conductivity ranks first in estimating the water quality index, followed by ammonia, and phosphate. The negatively ranked two attributes from feature selection such as boron and phenolpth alkalinity are considered as not important and removed from the dataset.

This feature selection method improved the river water quality dataset and finally the dataset with 10560 instances and 38 attributes has been developed and is named as WQI-SA dataset for reference.

3. WATER QUALITY INDEX PREDICTION MODEL

The problem of predicting the water quality index is formulated as a regression problem and solved using deep neural network architectures. Deep neural networks accurately describe, classify, and characterise data by using data inputs, weights, and biases. Deep neural networks have many layers of interconnected nodes, with two visible layers serving as input and output layers to improve prediction. Fig.3 illustrates the proposed framework architecture for the WQI prediction model. The pre-processed data is consumed by the deep learning model at the input layer, and the final prediction is made at the output layer. Large amounts of data can be used to train models, and the model improves as more data is added, as well as making high-quality predictions.

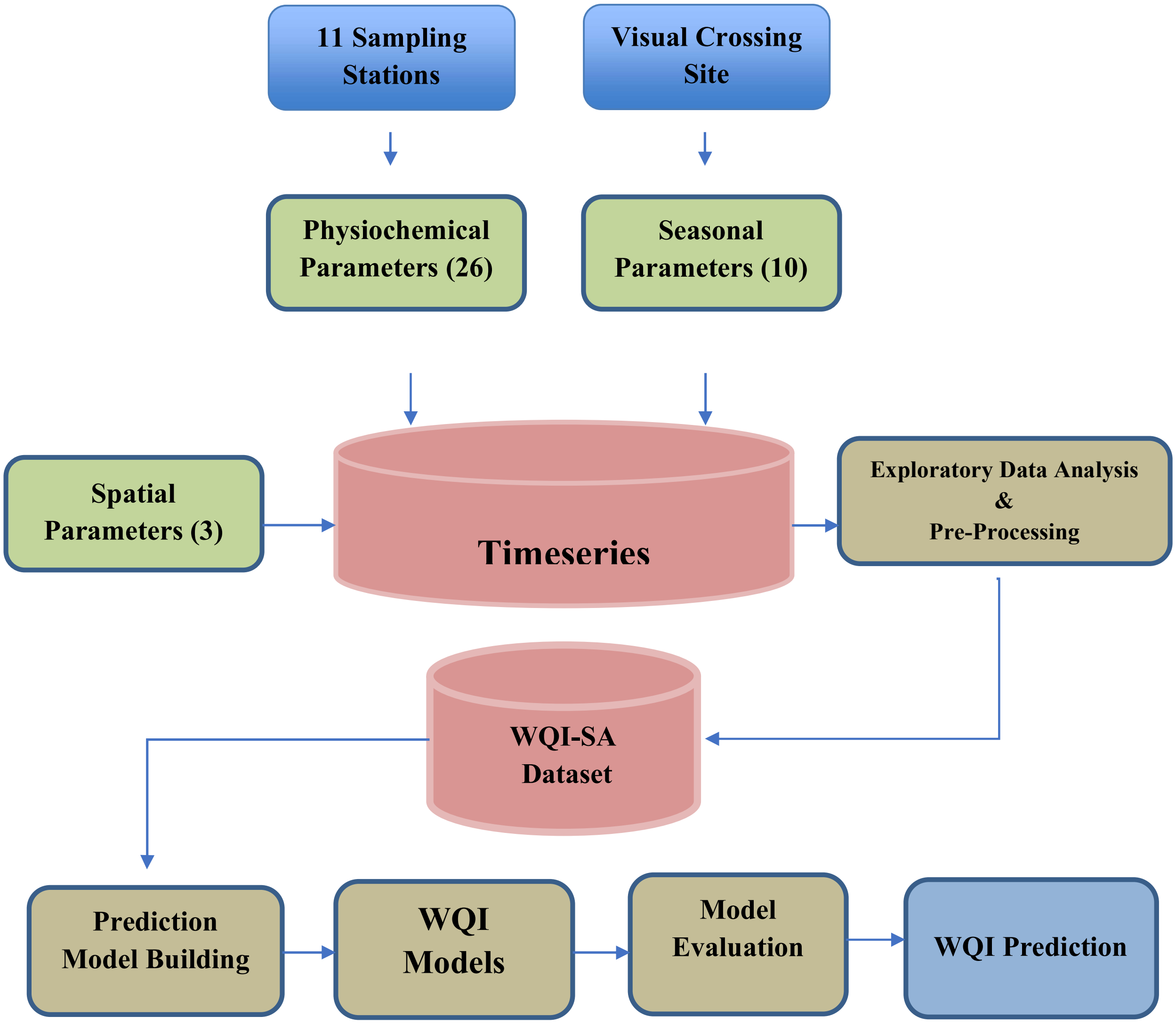


Fig. 3 Proposed WQI Model Architecture

Deep learning architectures such as recurrent neural networks, long short-term memory and gated recurrent networks are specifically designed and developed to train the sequence data and hence chosen in this work to build the river water quality index, prediction model. In a Recurrent Neural Network (RNN), the result from the previous section is used as input for the next. The Hidden state, which stores information about a sequence, is the primary and most crucial component of RNN. Due to their limited ability for long-term memory, RNNs are susceptible to the vanishing gradients problem. The primary problem for RNN is preserving data consistency across a large number of time steps. Gated recurrent networks and Long Short-term Memory are employed to overcome the vanishing gradient problem.

An LSTM recurrent unit seeks to recall all the earlier data encountered by the network and to forget irrelevant data. Each LSTM recurrent unit further stores a vector known as the Internal Cell State, which conceptually describes the information retained by the preceding LSTM recurrent unit. GRU employs a so-called update gate and reset gate to overcome the vanishing gradient problem of a typical RNN. Essentially, they are two vectors that determine what information is sent to the output. The unique characteristic of these systems is that they may be trained to retain knowledge from a long time ago without erasing it or removing extraneous data.

The 70% of instances of the WQI-SA dataset prepared as above are given as input to RNN and its variants LSTM and GRU for training the networks. The best hyperparameters are chosen during model training to make the model more effective at mapping the input features as independent variables to the target variable as the dependent variable.

Hidden layers, dense layers, optimizer, epoch, momentum, batch size, activation function, and dropout are examples of hyperparameters that are utilised in deep learning architectures to enhance model accuracy and fine-tune the WQI forecasting model. Hidden layers are the layers that are in between the input and output layers. A layer that is densely connected is one in which each layer receives input from all of the layers below it. The range is set between 5 and 10 units, and dense layers improve overall accuracy. Optimizers are methods that alter the properties of the neural network, like its weights and learning rate, to reduce losses and address optimization issues. The number of dataset complete iterations required is determined by the epoch size. Momentum is a unique hyperparameter that enables the search direction to be determined not only by the gradient from the current step but also by the gradients from previous steps. The model’s nonlinearity is introduced through activation functions. The activation function can split them into different layers and get a reduced output of the density layer. By passing randomly selected layers and limiting sensitivity to particular layer weights, the dropout layer helps prevent training overfitting. The speed at which a deep model replaces a previously learned concept with a new one is determined by the learning rate. Finally, the WQI prediction models are built by representation learning from the input instances using GRU, LSTM, and RNN with proper hyperparameters settings.

The effectiveness of the WQI forecasting model is evaluated using the evaluation metrics such as R2 score, root mean squared error, mean squared error, and mean absolute error. An estimator's mean squared error is the average of error squares or the difference between the predicted value and the actual value. The average difference between the predicted value and the actual value is what is used to calculate the mean absolute error. The root mean square error is used to measure a model's prediction error for quantitative data, which is a metric that indicates how well a regression line fits the data points. The R2 score value determines the accuracy of the model. If the R2 score value is high then the model is considered to be good in predicting the target variable and if the R2 score is less than 0.5 then the model is not considered to be good. 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 built with physicochemical and seasonal data is evaluated using the metrics with 30% of the dataset as the test set.

4. EXPERIMENT AND RESULTS

In our previous work, the experiments were carried out by training the time series WQI-PCA dataset that consists of only samples with physicochemical parameters and prediction models were built by employing deep neural architectures such as RNN, LSTM, and GRU. The prediction results of the models were obtained as shown in Table 4 and found that the GRU-WQI-based prediction model showed 88% of accuracy in predicting WQI.

Table 4: Prediction Result using WQI-PCA Dataset

Dataset	Model	MAE	MSE	RMSE	R2 Score
WQI-PCA	RNN-WQI	0.496	0.275	0.525	0.828
	LSTM-WQI	0.483	0.339	0.583	0.852
	GRU-WQI	**0.364**	**0.121**	**0.348**	**0.885**

In this work, using python libraries, the experiments were carried out by implementing the same three deep learning architectures. The RNN, LSTM, and GRU networks have been trained with the training dataset WQI-SA, which contains 7396 tagged samples and which is the 70% of the instances of the WQI-SA dataset. Evaluation of the prediction models is carried out to check the efficiency of the model using the metrics like R2 score, root mean squared error, mean squared error, and mean absolute error with the test data set containing 3170 instances.

The deep learning structures RNN, LSTM, and GRU are characterised by different hyperparameters, for example, dense layer values from 5 to 10 units, Optimizer as Adam optimizer. The epoch sizes were listed as 20, 50, 100, 150, and 200. The activation functions are defined as relu, sigmoid and tanh, and the momentum is set between 0.5 and 0.9. The dropout unit is 0.2, the learning rate is 0.1, and the batch size is set at either 32 or 64.

Table 5: Hyperparameters Setting for Deep Learning Architectures

Hyperparameter	Values	Hyperparameter	Values
Optimizer	Adam	Dropout	0.2,0.3
Dense Layer	5 to 10	Momentum	0.5 or 0.9
Epoch	20, 50, 100, 150, 200	Learning rate	0.1
Batch size	32/64	Activation function	Relu, sigmoid, and tanh

The evaluation results of RNN architecture with different epoch sizes such as 20, 50, 100, 150 and 200 are observed. The epoch size converges to 200 and from the prediction results, it is observed that the root mean squared error is 0.18. The mean absolute error is found as 0.15 when the epoch size 200 is set for the RNN algorithm. The evaluation results of RNN architectures found that the mean squared error is 0.0324 and the R2 score value is observed as 0.84.

The prediction results of the LSTM network for various epoch sizes such as 20,50, 100, 150, and 200 are noticed and found that epoch size 200 is showing better prediction results. The mean absolute error is 0.046 and the mean squared error is 0.01 for the LSTM algorithm when the epoch size is set to 200 and the optimizer is Adam. Similarly, it is found that the root mean squared error shown by LSTM-based WQI prediction model is 0.1 and the R2 score value is found to be 0.9 when the epoch size is set to 200.

The results of GRU based WQI prediction model for various epoch sizes 20, 50,100,150 and 200 are measured. The mean absolute error is found to be 0.134 and the mean squared error of 0.0576 for epoch size 200 is observed. It is observed that the root mean squared error is 0.24 and the R2 score is 0.88 found for the GRU-based WQI forecasting model when the epoch size is 200.

The performance evaluation of the WQI prediction model based on the WQI-SA dataset and deep learning architectures RNN, LSTM and GRU concerning various metrics such as mean squared error, mean absolute error, root mean squared error and R2 score using different epochs is shown in Table 6a, 6b and 6c and illustrated in Fig 4a, Fig 4b Fig 4c and Fig 4d.

Table 6a. Prediction Results of RNN Models for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	200	0.15	0.0324	0.18	0.84
	150	0.19	0.0484	0.22	0.83
	100	0.25	0.0625	0.25	0.832
	50	0.28	0.0784	0.28	0.824
	20	0.39	0.0961	0.31	0.79

Table 6b. Prediction Results of LSTM Models for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	200	0.046	0.01	0.1	0.9
	150	0.091	0.0169	0.13	0.89
	100	0.13	0.0324	0.18	0.886
	50	0.135	0.057	0.24	0.884
	20	0.142	0.084	0.29	0.87

Table 6c. Prediction Results of GRU Models for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	200	0.134	0.0576	0.24	0.88
	150	0.146	0.0625	0.25	0.873
	100	0.15	0.0852	0.292	0.877
	50	0.17	0.102	0.32	0.867
	20	0.19	0.122	0.35	0.85

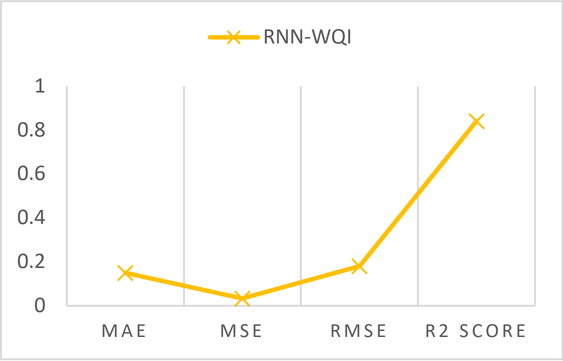


Fig. 4a. Evaluation of RNN Architecture

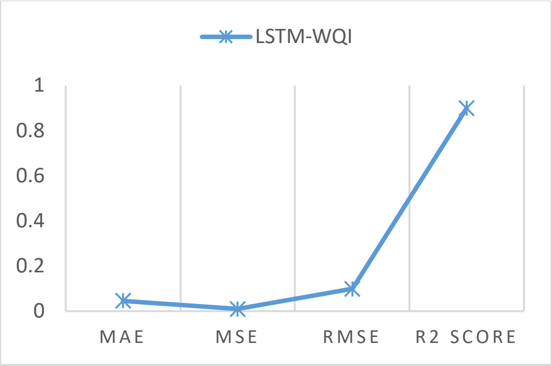


Fig. 4b. Evaluation of LSTM Architecture

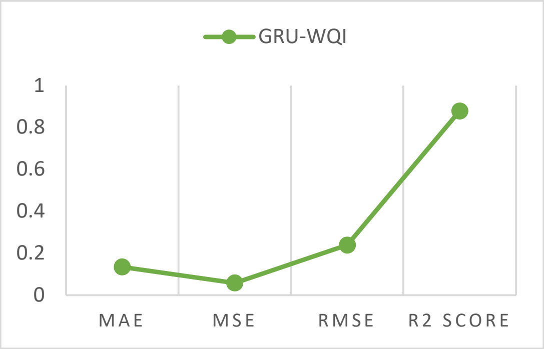


Fig. 4c. Evaluation of GRU Architecture

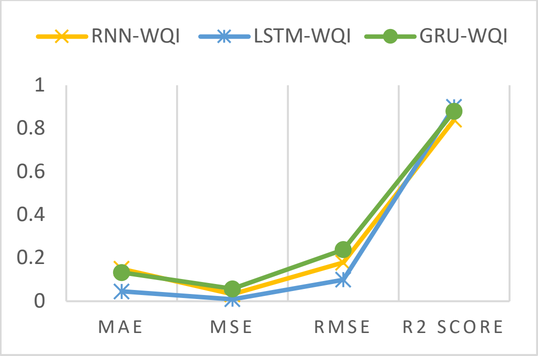


Fig 4d. Performance Analysis using WQI-SA Dataset

Various experiments have been carried out with different dropout rates such as 0.2 and 0.3 for building WQI prediction models using the WQI-SA dataset and the experimental results concerning the same evaluation metrics are shown in Table 7a, 7b and 7c.

Table 7a. Prediction Results of RNN for WQI-SA Dataset and Different Dropouts

Dataset	Dropout	MAE	MSE	RMSE	R2 Score
WQI-SA	0.3	0.15	0.18	0.424	0.84
	0.2	0.17	0.2	0.44	0.83

Table 7b. Prediction Results of LSTM for WQI-SA Dataset and Different Dropouts

Dataset	Dropout	MAE	MSE	RMSE	R2 Score
WQI-SA	0.3	0.046	0.01	0.1	0.9
	0.2	0.091	0.0169	0.13	0.89

Table 7c. Prediction Results of GRU for WQI-SA Dataset and Different Dropouts

Dataset	Dropout	MAE	MSE	RMSE	R2 Score
WQI-SA	0.3	0.134	0.0576	0.24	0.88
	0.2	0.146	0.0625	0.25	0.873

The prediction results of WQI models for various epochs and dropouts have been observed while implementing deep learning algorithms to discover the best prediction results. It is proved that the models trained with 200 epochs and dropout rate 0.3 with other hyperparameters Adam optimizer, the learning rate is 0.1, set for RNN, LSTM and GRU produced best results and are shown in Table 8.

Table 8. Best Prediction Results of Deep Learning Algorithms

Dataset	Dropout	Epoch	Algorithm	MAE	MSE	RMSE	R2 Score
WQI-SA	0.3	200	RNN	0.15	0.0624	0.25	0.84
			LSTM	0.046	0.01	0.1	0.9
			GRU	0.134	0.0576	0.24	0.88

From the above results, it is observed that the LSTM-based WQI prediction model shows promising results with a high R2 score value and less error rate. The mean absolute error for LSTM based forecasting model is found less as compared to RNN and GRU algorithms. The root mean squared error is observed to be less for LSTM architecture when compared with RNN and GRU-based prediction model results. The R2 score value defines the accuracy of the model and is observed to be high for LSTM-based forecasting models compared with other prediction models.

The performance of these forecasting models is compared with the previously developed WQI-PCA dataset for WQI prediction models. The evaluation results of RNN architectures are observed to mean absolute error value is found as 0.496 when the WQI-PCA dataset is employed whereas when WQI-SA is given the result is observed as 0.15. The root mean squared error is observed as 0.525 while employing the WQI-PCA dataset and 0.18 error when using the WQI-SA dataset for evaluation results of RNN. The R2 score value is observed as 0.82 while using WQI-PCA dataset and it is 0.84 when employing the WQI-SA dataset with the RNN algorithm.

Similarly, the experimental results of LSTM-based WQI prediction models trained with two datasets WQI-PCA and WQI-SA are examined. It is observed that the mean absolute error is 0.483 for the WQI-PCA dataset and 0.046 when the WQI-SA dataset is used. The root mean squared error is observed as 0.583 while employing the WQI-PCA dataset and a 0.1 error rate when using the WQI-SA dataset for WQI prediction. The R2 score value is observed as 0.852 while using the WQI-PCA dataset and it is 0.9 when employing the WQI-SA dataset for building the LSTM-based WQI prediction model.

It is observed that the prediction results of GRU-based WQI models show different error rates with different datasets. The mean absolute error is found to be 0.364 when WQI-PCA is employed and when WQI-SA is used it gives a 0.134 error rate. Similarly, the root mean squared error is found to be 0.121 with the WQI-PCA dataset and 0.0576 with the WQI-SA dataset. The R2 score value is found to be 0.88 when the WQI-PCA dataset is trained and the WQI-SA dataset is trained in the GRU network, the R2 score is obtained as 0.88.

From the performance results, it is observed that the GRU-based prediction model using WQI-PCA dataset gives less error rate and a high R2 score value. The LSTM-based WQI prediction model using the WQI-SA dataset shows high R2 score values and fewer error rates with metrics like mean absolute error, mean squared error and root mean squared error.

The comparative performance of WQI prediction models based on two different datasets such as WQI-PCA and WQI-SA is shown in Table 9 and illustrated in Fig. 5.

Table 9. Comparative Performance of Models Based on Two Datasets

Dataset	Algorithm	MAE	MSE	RMSE	R2 Score
WQI-PCA	RNN-WQI	0.496	0.275	0.525	0.828
	LSTM-WQI	0.483	0.339	0.583	0.852
	GRU-WQI	**0.364**	**0.121**	**0.348**	**0.885**
WQI-SA	RNN-WQI	0.15	0.0324	0.18	0.84
	LSTM-WQI	**0.046**	**0.01**	**0.1**	**0.9**
	GRU-WQI	0.134	0.0576	0.24	0.88

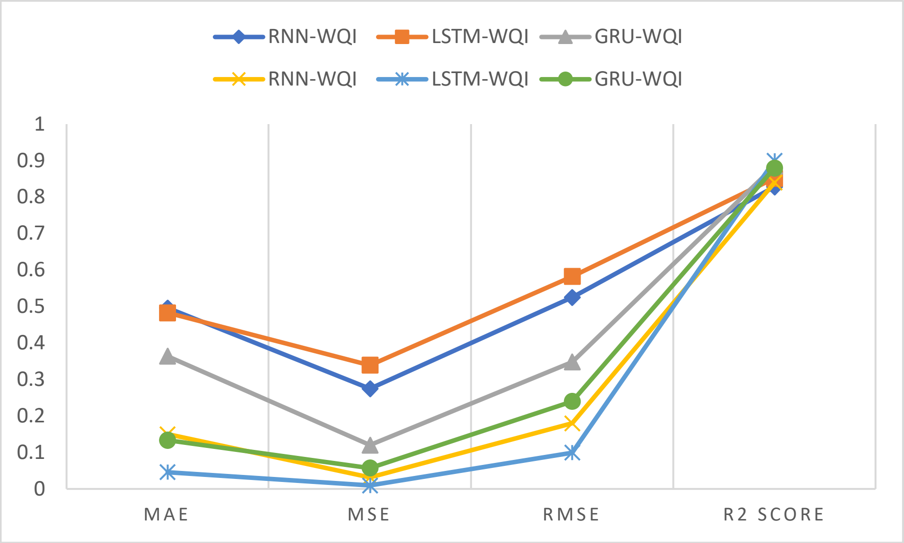


Fig 5. Overall Comparative Analysis of WQI Prediction Models

The investigations made in this research proved that the machine learning approach is useful for developing predictive models like water quality index prediction. It is confirmed that the recent deep learning approach improves the prediction accuracy of different WQI predictive models. Through feature selection, the association between the pool of predictors and the targeted variable is strengthened which enables deep neural network architectures GRU, LSTM, and RNN to improve the learning of trends in the data. The prediction rate of WQI models is increased through learning the self-extracted features in GRU, LSTM, and RNN networks. The error rate of trained models is decreased by properly configuring the hyperparameters during network training. The addition of seasonal parameters in the time series data enhances the quality of WQI prediction as they are more influential in water quality determination. The enhanced water quality prediction model with seasonal time series data has proven to be an effective tool in predicting water quality in different locations.

5. Conclusion

This study demonstrated the importance of seasonal data in building WQI prediction models. The application of deep learning architectures for river water quality time series forecasting was attempted to prove that deep learning is an effective approach for accurate WQI prediction. The seasonal data collected from the visual crossing site during the period 2016 to 2020 were pooled with the physiochemical parameters of river water collected from the Bhavani River and a new time series dataset was developed. The river water quality forecasting model has been designed and developed using deep learning architectures such as LSTM, RNN and GRU. The performance of the new models was evaluated and compared with the prediction results of models trained with only physiochemical parameters. From the evaluation results, it is observed that the augmentation of seasonal data enhanced the efficiency of the water quality prediction model. A generalized model has been developed, which can be used in predicting the water quality of any river. The developed model can even be used as the pre-trained model for applying transfer learning.

References

1. Gopal Krishan, Surjeet Singh, Kumar CP, Suman Gurjar and Ghosh NC, (2016), "Assessment of Water Quality Index (WQI) of Groundwater in Rajkot District, Gujarat, India", Journal of Earth Science & Climatic Change, Volume 7, Issue 3, 1000341 (2016)

2. Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto (2019), 'Efficient Water Quality Prediction Using Supervised Machine Learning'- Journal of Water.

3. M. Ezhilarasi and V. Senthilkumar (2018), Geo-Chemical Analysis for Groundwater Quality Using Geospatial Application, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04.

4. Shuangyin Liu, Haijiang Taia, Qisheng Dinga, Daoliang Lia, Longqin Xub, Yaoguang Weia (2011), 'A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction'- Journal of Elsevier.

5. Salisu Yusuf Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz and Azrul Amri Jamal (2015), 'Classification Model for Water Quality using Machine Learning Techniques' - International Journal of Software Engineering and Its Applications Vol. 9, No. 6 , pp. 45-52.

6. 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

7. 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

8. 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

9. J. P. Nair and M. S. Vijaya, "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), 2021, pp. 1747-1753.

10. Nair, Jitha P., and M. S. Vijaya. 'River Water Quality Prediction and Index Classification Using Machine Learning'. *Journal of Physics: Conference Series*, vol. 2325, no. 1, Aug. 2022, p. 012011.

11. Heddam, S., 2014. Generalized regression neural network-based approach for modelling hourly dissolved oxygen concentration in the Upper Klamath River, Oregon, USA. Environmental Technology (United Kingdom) 35, 1650–1657. 10.1080/09593330.2013.878396.

12. Nair, J.P., Vijaya, M.S. (2023). Exploratory Data Analysis of Bhavani River Water Quality Index Data. In: Kumar, S., Hiranwal, S., Purohit, S.D., Prasad, M. (eds) Proceedings of International Conference on Communication and Computational Technologies. Algorithms for Intelligent Systems. Springer, Singapore.
13. 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.
14. Basant, N., Gupta, S., Malik, A., Singh, K.P., 2010. Linear and nonlinear modelling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water -A case study. Chemometr.Intellig.Lab.Syst.104,172–180.

15. Heddam, S., Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 559, 499–509.

16. Li, G., 2006. Stream temperature and dissolved oxygen modelling in the Lower Flint River Basin. PhD Dissertation. University of Georgia, Athens, GA.

17. Wang, Y., 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) (pp. 1-5). IEEE.

18. 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

19. Stroomberg G. J., Freriks I. L., Smedes F. and Co®no W. P. (1995) In Quality Assurance in Environmental Monitoring, ed. P. Quevauviller. VCH, Weinheim.

20. G Tan, J Yan, C Gao, and S Yang, Prediction of water quality time series data based on least squares support vector machine, Procedia Engineering, Vol. 31, 2012, pp. 1194-1199.

21. WC Leong, A Bahadori, J Zhang, and Z Ahmad, Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), International Journal of River Basin Management, Vol. 19, 2021, pp. 149-156.

22. Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H. & Maashi, M. Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics 2020, 6659314 (2020).

23. Yang, Y. et al. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environmental Science and Pollution Research (2021) .

24. Kamaraj M, Rangarajan S (2022) Predicting the future land use and land cover changes for Bhavani basin, Tamil Nadu, India, using QGIS MOLUSCE plugin. Environ Sci Pollut Res (2022).

25. Arunkumar R, Thambusamy, Velmurugan (2021) An exploratory data analysis process on groundwater quality data. 54:41–48

26. Haghiabi, Amir Hamzeh, et al. 'Water Quality Prediction Using Machine Learning Methods'. Water Quality Research Journal, vol. 53, no. 1, Feb. 2018, pp. 3–13.

27. Ubah, J.I., Orakwe, L.C., Ogbu, K.N. et al. Forecasting water quality parameters using artificial neural network for irrigation purposes. Sci Rep 11, 24438 (2021).

28. Kargar, K. et al. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. Eng. Appl. Comput. Fluid Mech. 14, 311–322 (2020).

29. Alizadeh, M. J. et al. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. Eng. Appl. Comput. Fluid Mech. 12(1), 810–823 (2018).

30. WHO 2019 Drinking-water
https://www.who.int/news-room/factsheets/detail/drinking-water