5. DEEP LEARNING FOR WQI PREDICTION MODELS WITH POOLED PARAMETERS

Water Quality Index prediction models are tools used to evaluate and predict the quality of water based on various physical, chemical, and biological parameters. WQI models aim to provide a comprehensive and objective assessment of water quality by aggregating several parameters into a single index called WQI. Deep learning models have emerged as a promising alternative for WQI prediction using various data sources such as physiochemical and seasonal data. 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. In this work, seasonal features are considered along with physiochemical parameters, for trend analysis, and to construct an improved water quality prediction model using RNN and its variants.

WQI PREDICTION MODEL USING POOLED FEATURES AND RNN VARIANTS

This work aims to build a predictive model for WQI by utilizing deep-learning architecture, specifically RNN and its variants. The model is designed to capture and learn patterns present in the time series data, which comprises both physiochemical and seasonal parameters. The dynamics of river water quality are influenced by physiochemical, seasonal parameters, which experience abrupt changes due to climatic variations. Extensive research in the literature has revealed the significant impact of seasonal parameters on the water quality index and its predictive capabilities when analysed over time series data. The relationship between simultaneous rainfall and humidity is strong, as the evaporation of rainwater enhances relative humidity. Building on this observation, the study aims to leverage advanced deep-learning techniques to enhance the accuracy and effectiveness of WQI prediction model.

Methodology

RNN, LSTM and GRU are highly valuable architectures in sequence and time series data analysis and training, due to their ability to effectively capture temporal dependencies and patterns. These architectures, with their recurrent nature and memory mechanisms, excel in recognizing trends and dependencies in time series data. The methodology of the proposed WQI prediction model consists of important tasks which include 1. data collection 2. EDA and dataset creation 3.

construction of WQI prediction model 4. model evaluation. The framework of the WQI prediction model based on the pooled parameters is depicted in Fig.5.1.

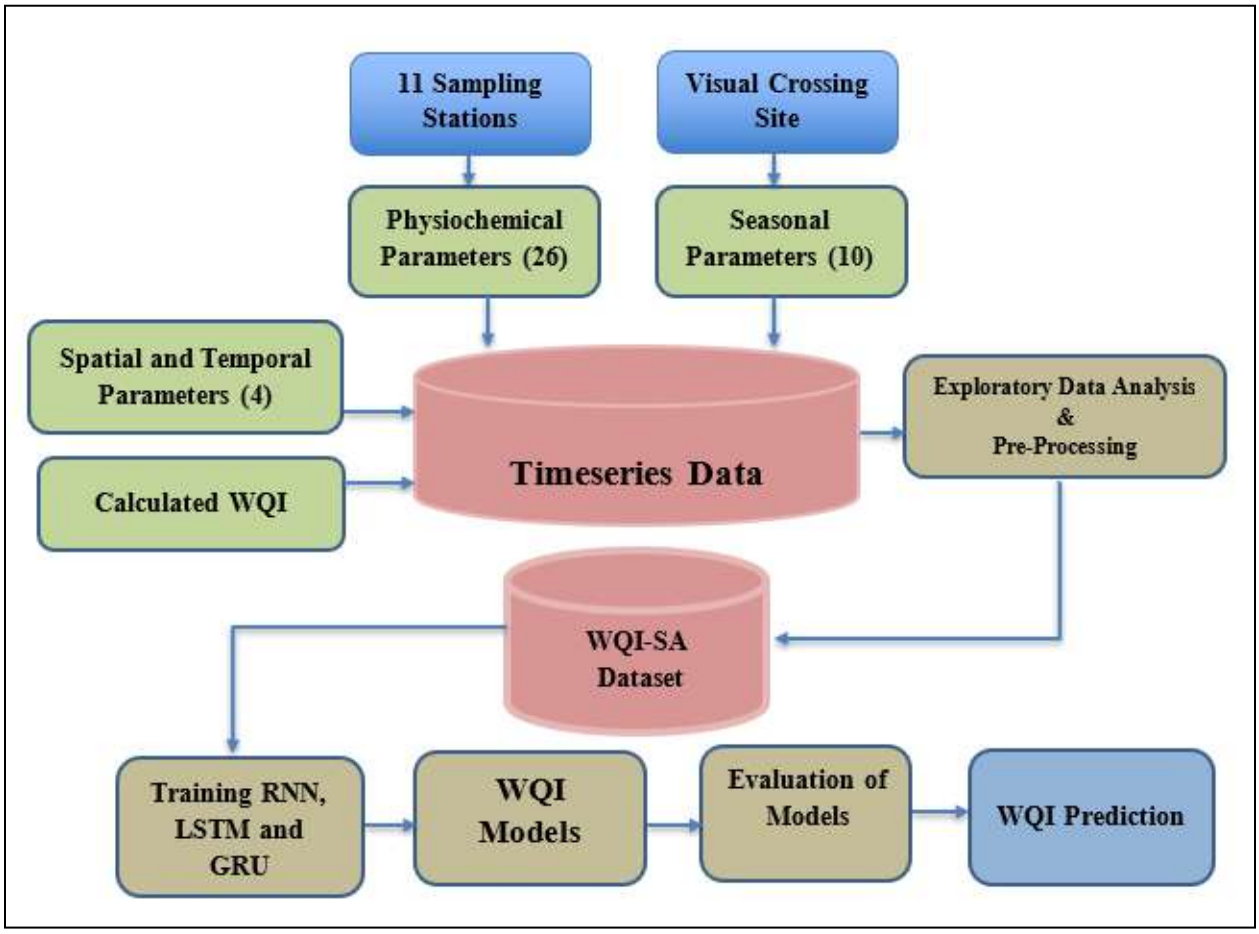

Fig. 5.1. Framework of the WQI Prediction Model Based on Pooled Parameters and RNN Variants

Data Collection and Dataset Preparation

The 26 different physicochemical parameters such as pH, conductivity, turbidity, phenolpth alkalinity, total alkalinity, chloride, chemical oxygen demand, total Kjeldahl nitrogen, ammonia, hardness, Ca.hardness, Mg. hardness, sulphate, sodium, total suspended solids, total dissolved solids, fixed dissolved solids, phosphate, boron, potassium, biological oxygen demand, fluoride, nitrate, dissolved oxygen, total coliform and faecal coliform, are collected from the monitoring stations across Bhavani River. The seasonal characteristics such as temperature, dew, humidity, sea level pressure, precipitation, precip over, wind speed, wind direction, cloud cover, and visibility are collected from visual crossing sites for the corresponding locations of monitoring

stations of the Bhavani River. The water quality index value for each sample is calculated and assigned to the corresponding samples as a target variable. A time series data with 10560 and 41 attributes including 26 physiochemical parameters, 10 seasonal parameters, longitude, latitude, station ID, date and calculated WQI, has been created.

The river water quality data is subjected to EDA to understand its characteristics and assess the importance of each parameter in determining the WQI. Various statistical techniques such as heatmap analysis, boxplot analysis, pair plot analysis, and histogram analysis are employed to study and comprehend the distribution of parameter values. The min-max normalization is applied to water quality to standardise the parameter values. The select K best feature selection method is used to remove irrelevant attribute and to substantially improved the river water quality dataset. Finally, 10560 tagged instances with 38 attributes are developed and referred to as the WQI-SA dataset as mentioned in Table XII of Chapter 3.

Model Building

Deep learning architectures such as RNN, LSTM, and GRU 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 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. 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. The unique characteristic of GRU is that they may be trained to retain knowledge from a long time ago without erasing it or removing extraneous data. During training, these architectures optimize their parameters using backpropagation through time, adjusting weights to minimize the error between the predicted and actual outputs, thereby enabling them to learn complex temporal patterns in the data.

The 80% of instances of the WQI-SA dataset are given as input to RNN and its variants LSTM and GRU for training the networks independently. The best hyperparameters are chosen during model training to make the model more effective 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 some of the hyperparameters that are utilized in deep learning architectures to enhance model accuracy and fine-tune the 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 datasets 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 learning rate determines the speed at which a deep model replaces a previously learned concept with a new one. Finally, three independent WQI prediction models are built by learning water quality patterns from the input instances of the WQI-SA dataset through training RNN, LSTM and GRU with proper hyperparameters settings. These models are called as RNN-WQI-SA, LSTM-WQI-SA and GRU-WQI-SA models for reference. The effectiveness of the WQI forecasting models is evaluated using MAE, MSE, RMSE and R2 score.

Experiments and Results

The experiments have been carried out by implementing deep learning algorithms such as RNN, LSTM and GRU and by training the Bhavani River water dataset WQI-SA using Python libraries under TensorFlow, Keras and scikit learn. The training dataset contains 8124 tagged instances of the WQI-SA dataset. The evaluation of the prediction models is carried out using the metrics such as MAE, MSE, RMSE and R2 score values with the test data set containing 2009 tagged instances of the WQI-SA dataset.

The deep learning algorithms RNN, LSTM, and GRU are implemented by defining hyperparameters, dense layer values from 5 to 10 units, and optimizer as adam optimizer. The epoch sizes were listed as 20, 50, 100, 150, 200 and 500. The activation function used to train the

model is relu 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. From the experimental results, it is proved that with momentum as 0.8, epoch sizes 500, drop out 0.3 and with relu activation function better results are achieved. The hyperparameter settings for training deep neural networks are tabulated in Table XXI.

Table XXI. Hyperparameters Setting for Training Deep Neural Networks

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

The results of the RNN-based WQI prediction model (RNN-WQI-PCA model) are experimented with various epochs such as from 20 to 500 where various metrics are measured at different epochs. At epoch 500, the RNN model achieves an MAE of 0.424, indicating the average absolute difference between the predicted and actual values. The MSE is calculated as 0.384, representing the average of squared differences. The RMSE is 0.6196, which is the square root of the MSE. The R2 score, measuring the goodness of fit, is 0.82, indicating a high level of prediction accuracy. Moving to epoch 200, the MAE increases slightly to 0.459, while the MSE becomes 0.392. The RMSE is 0.6260, and the R2 score remains relatively high at 0.813. As the number of epochs decreases, the MAE and MSE values gradually increase, indicating a larger difference between the predicted and actual values.

At epoch 150, the MAE is 0.482, and the MSE is 0.424, resulting in an RMSE of 0.6511. The R2 score decreases to 0.806, suggesting a slightly lower level of prediction accuracy compared to the previous epochs. At epoch 100, the MAE increases further to 0.512, and the MSE becomes 0.462. The RMSE is 0.6797, and the R2 score remains relatively stable at 0.80. With only 50 epochs, the MAE reaches 0.537, and the MSE increases to 0.527. The RMSE becomes 0.7259, while the R2 score decreases slightly to 0.79. Finally, at epoch 20, the MAE is 0.579, the MSE is

0.561, and the RMSE is 0.7489. The R2 score drops to 0.78. These values reflect the performance of the RN-WQI-SA model on the WQI-SA dataset at different epochs, providing insight into the prediction results which are tabulated in Table XXII.

Table XXII. Prediction Results of RNN-WQI-SA Model for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	500	0.428	0.384	0.6196	0.82
	200	0.459	0.392	0.6260	0.813
	150	0.482	0.424	0.6511	0.806
	100	0.512	0.462	0.6797	0.8
	50	0.537	0.527	0.7259	0.79
	20	0.579	0.561	0.7489	0.78

The prediction results of the LSTM-based WQI prediction model (LSTM-WQI-SA model) for different epochs on the WQI-SA dataset. At epoch 500, the LSTM-WQI-SA model achieves an MAE of 0.298, indicating the average absolute difference between the predicted and actual values. The MSE is calculated as 0.2084, representing the average of squared differences. The RMSE is 0.4565, which is the square root of the MSE. The R2 score, measuring the goodness of fit, is 0.856, indicating a high level of prediction accuracy. Moving to epoch 200, the MAE increases slightly to 0.304, while the MSE becomes 0.239. The RMSE is 0.4888, and the R2 score remains relatively high at 0.85. As the number of epochs decreases, the MAE and MSE values gradually increase, indicating a larger difference between the predicted and actual values. At epoch 150, the MAE is 0.328, and the MSE is 0.274, resulting in an RMSE of 0.5234. The R2 score decreases to 0.843, suggesting a slightly lower level of prediction accuracy compared to the previous epochs.

At epoch 100, the MAE increases further to 0.371, and the MSE becomes 0.291. The RMSE is 0.5394, and the R2 score remains relatively stable at 0.839. With only 50 epochs, the MAE reaches 0.398, and the MSE increases to 0.328. The RMSE becomes 0.5727, while the R2 score decreases slightly to 0.83. Finally, at epoch 20, the MAE is 0.402, the MSE is 0.367, and the RMSE is 0.6058. The R2 score drops to 0.827. These values illustrate the performance results of

the LSTM-WQI-SA model on the WQI-SA dataset at different epochs, providing insight into the prediction results which are tabulated in Table XXIII.

Table XXIII. Prediction Results of LSTM-WQI-SA Model for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	500	0.298	0.2084	0.4565	0.856
	200	0.304	0.239	0.4888	0.85
	150	0.328	0.274	0.5234	0.843
	100	0.371	0.291	0.5394	0.839
	50	0.398	0.328	0.5727	0.83
	20	0.402	0.367	0.6058	0.827

The prediction results of the GRU-based WQI prediction model (GRU-WQI-SA model) for different epochs on the WQI-SA dataset. At epoch 500, the GRU-WQI-SA model achieves an MAE of 0.39, indicating the average absolute difference between the predicted and actual values. The MSE is calculated as 0.2149, representing the average of squared differences. The RMSE is 0.4636, which is the square root of the MSE. The R2 score, measuring the goodness of fit, is 0.839, indicating a relatively high level of prediction accuracy. Moving to epoch 200, the MAE increases slightly to 0.412, while the MSE becomes 0.2342. The RMSE is 0.4839, and the R2 score decreases to 0.83. As the number of epochs decreases, the MAE and MSE values gradually increase, indicating a larger difference between the predicted and actual values.

At epoch 150, the MAE is 0.436, and the MSE is 0.269, resulting in an RMSE of 0.5187. The R2 score decreases to 0.823, suggesting a slightly lower level of prediction accuracy compared to the previous epochs. At epoch 100, the MAE increases further to 0.452, and the MSE becomes 0.287. The RMSE is 0.5357, and the R2 score remains relatively stable at 0.82. With only 50 epochs, the MAE reaches 0.462, and the MSE increases to 0.315. The RMSE becomes 0.5612, while the R2 score decreases slightly to 0.803. Finally, at epoch 20, the MAE is 0.474, the MSE is 0.348, and the RMSE is 0.5899. The R2 score drops to 0.793. These values highlight the performance of the GRU model on the WQI-SA dataset at different epochs, providing insight into the prediction results which are tabulated in Table XXIV.

Table XXIV. Prediction Results of GRU-WQI-SA Model for Various Epochs

Dataset	Epochs	MAE	MSE	RMSE	R2 Score
WQI-SA	500	0.39	0.2149	0.4636	0.839
	200	0.412	0.2342	0.4839	0.83
	150	0.436	0.269	0.5187	0.823
	100	0.452	0.287	0.5357	0.82
	50	0.462	0.315	0.5612	0.803
	20	0.474	0.348	0.5899	0.793

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 XXV.

Table XXV. Results of WQI Prediction Models for Different Dropout Rates

Dataset	Algorithm	Dropout	MAE	MSE	RMSE	R2 Score
WQI-SA	RNN	0.3	0.428	0.384	0.6197	0.82
		0.2	0.482	0.424	0.6512	0.806
	LSTM	0.3	0.298	0.2084	0.4565	0.856
		0.2	0.328	0.274	0.5235	0.843
	GRU	0.3	0.39	0.2149	0.4636	0.839
		0.2	0.436	0.269	0.5187	0.823

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 500 epochs and dropout rate 0.3 with other hyperparameters like adam optimizer, momentum as 0.8 and activation function as relu for RNN, LSTM and GRU produced the best results and are shown in Table XXVI and depicted in Fig. 5.2.

Table XXVI. Prediction Results of all Three WQI Models

Dataset	Dropout	Epoch	Models	MAE	MSE	RMSE	R2 Score
WQI-SA	0.3	500	RNN-WQI-SA	0.428	0.384	0.6197	0.82
			LSTM-WQI-SA	**0.298**	**0.2084**	**0.4565**	**0.856**
			GRU-WQI-SA	0.39	0.2149	0.4636	0.839

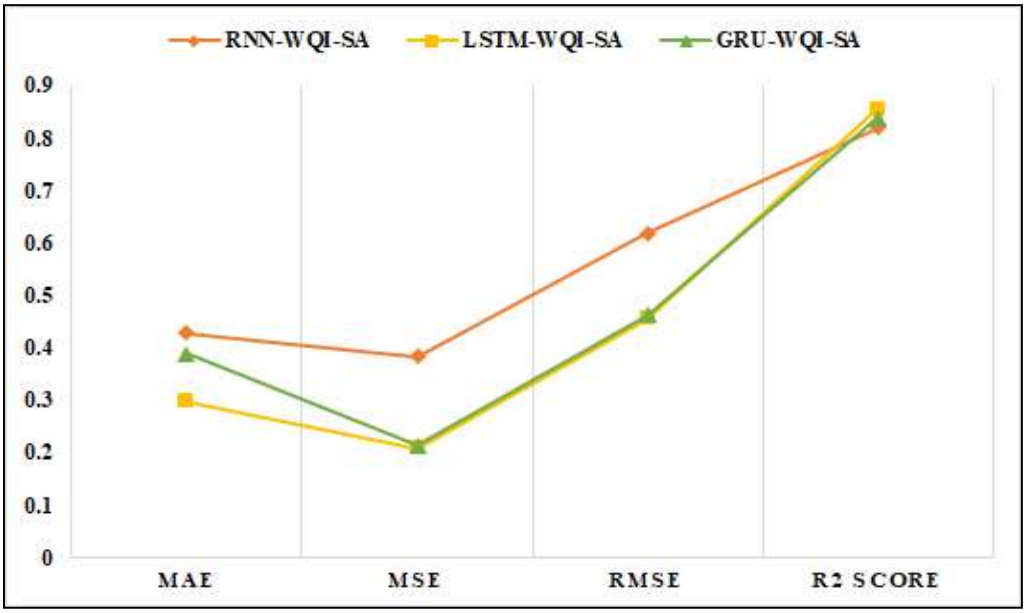

Fig.5.2. Prediction Performance of all Three WQI Models

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 the LSTM-WQI-SA model when compared with RNN-WQI-SA and GRU-WQI-SA prediction model results. The R2 score value defines the accuracy of the model and is observed to be high for the LSTM-WQI-SA forecasting model compared with other prediction models.

Comparative Analysis WQI Models based on WQI-PCA and WQI-SA Datasets

The performance results of prediction models built using two distinct datasets such as WQI-PCA and WQI-SA are compared to analyse influence of seasonal parameters the efficiency

of the prediction models. From the comparative study, it is evident that the prediction models built using the WQI-SA dataset performed better than the models built using the WQI-PCA dataset. The LSTM-WQI-SA model emerged as the most accurate one, exhibiting the lowest MAE, MSE, and RMSE, along with the highest R2 Score. Here it is evident that the incorporation of seasonal parameters has improved the efficacy of the WQI prediction models. The performance analysis of the WQI prediction models is tabulated in Table XXVII and illustrated in Fig. 5.3.

Table XXVII. Performance Comparison of WQI Models based on WQI-PCA and WQI-SA Datasets

Dataset	Models	MAE	MSE	RMSE	R2 Score
WQI-PCA	RNN-WQI-PCA	0.512	0.408	0.6387	0.8
	LSTM-WQI-PCA	0.393	0.2401	0.4900	0.838
	GRU-WQI-PCA	0.364	0.2098	0.4580	0.845
WQI-SA	RNN-WQI-SA	0.428	0.384	0.6197	0.82
	LSTM-WQI-SA	**0.298**	**0.2084**	**0.4565**	**0.856**
	GRU-WQI-SA	0.39	0.2149	0.4636	0.839

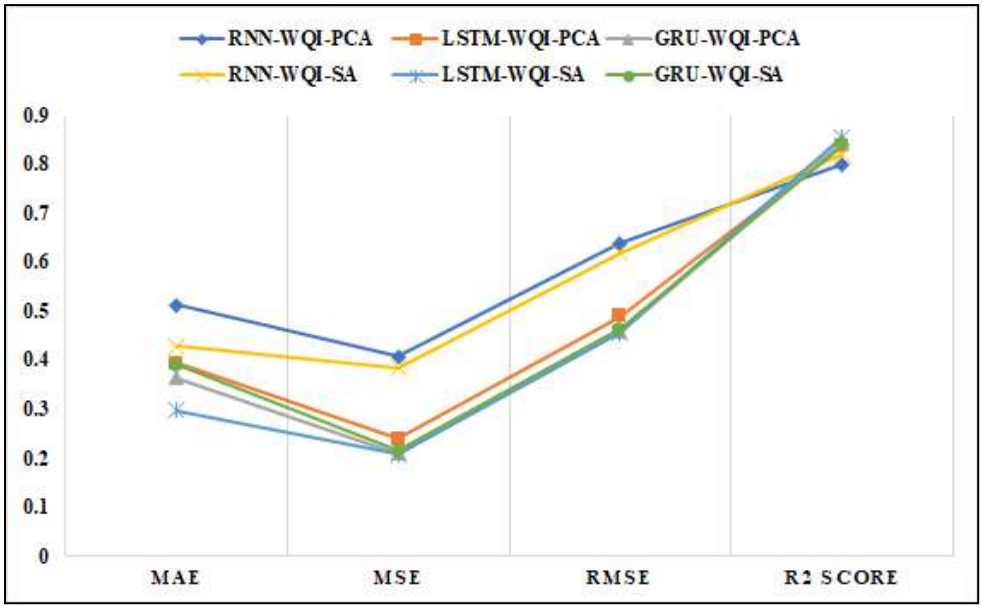

Fig. 5.3. Performance Comparison of WQI Models based on WQI-PCA and WQI-SA Datasets

Findings

The investigations made in this work proved that the deep learning approach is useful for developing predictive models for WQI prediction using time series data. 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. Through adding seasonal parameters, the association between the pool of predictors and the targeted variable is strengthened which enables deep neural network algorithms RNN, LSTM and GRU to improve the learning of trends in the data. The prediction rate of WQI models is increased through learning the self-extracted features in RNN, LSTM, and GRU networks. The error rate of trained models is decreased by properly configuring the hyperparameters during network training. The enhanced water quality prediction model with seasonal time series data has proven to be an effective tool in predicting water quality.

SUMMARY

This chapter described the construction of an improved water quality prediction model using pooled parameters and RNN and its variants. The implementation of various deep-learning techniques for building WQI models has been described in detail. Three independent models have been built using RNN variants and the performance results have been reported. The impact of seasonal parameters in determining WQI is analysed through this work. The inclusion of seasonal parameters in the time series data elevates the accuracy of WQI prediction, owing to their greater influence on water quality prediction. The construction of the WQI prediction model with more sophisticated architecture, temporal fusion transformer, will be explained in the following chapter.

Remarks

The paper titled "Enhanced Water Quality Prediction Model with Seasonal Time Series Data" has been published in the Journal of Data Acquisition and Processing, Vol.38(1), 2023.pp 1283-1303. ***(Scopus Indexed)***