Deep Learning for Water Quality Index Prediction in Peripheral rivers of Dhaka City

*

1st Parvez Ahammed
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
parvez.cse.200104129@aust.edu

2nd Tabssum Tara Lamia
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
tabassum.cse.200104128@aust.edu

3rd A. M. Sajid Hasan
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
hasan.cse.200104148@aust.edu

4th Ashraf Uz Zaman Shuvo
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
ashraf.cse.200104144@aust.edu

***Abstract*—This research employs advanced time series forecasting techniques to analyze the Water Quality Index (WQI) in the Peripheral rivers of Dhaka City. Utilizing a comprehensive dataset from environmental monitoring stations, the study explores temporal patterns over a twenty-year period (2003-2023). The research aims to provide insights into the evolving dynamics of water quality, offering valuable information for proactive environmental management in the face of urbanization challenges.**

***Index Terms*—water quality prediction; Deep Learning; LSTM; GRU;**

I. Introduction

As urbanization accelerates globally, the management of water resources becomes increasingly imperative, particularly in densely populated areas such as Dhaka City. This research endeavors to address the intricate dynamics of water quality in the Peripheral rivers encircling Dhaka by focusing on the Time Series Forecasting of the Water Quality Index (WQI). The Water Quality Index serves as a comprehensive metric, encapsulating diverse parameters that collectively define the health of aquatic ecosystems.

This paper concentrates on the application of advanced time series forecasting techniques to unravel the temporal evolution of the Water Quality Index in Dhaka City's Peripheral rivers. By embracing a chronological perspective, we aim to discern patterns, trends, and anticipate future values, thereby enhancing our comprehension of the nuanced interplay of factors influencing water quality.

The urgency of forecasting the Water Quality Index stems from the escalating anthropogenic pressures on water bodies due to urban expansion and industrialization. This research seeks to elucidate the temporal dynamics governing water quality, providing a foundation for proactive environmental management. By forecasting the WQI, we aspire to furnish decision-makers with insights essential for formulating policies that safeguard both the ecological integrity of the rivers and the well-being of communities reliant on these water sources.

Our methodology incorporates advanced time series forecasting models, specifically the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) algorithms. These models, known for their capability to capture temporal dependencies, are employed to predict water quality indices in Dhaka City's Peripheral rivers. Utilizing a rich dataset from strategically positioned monitoring stations, we apply LSTM and GRU to comprehensively model and forecast the nuanced temporal patterns in water quality. This dual approach ensures a holistic understanding of the dynamic factors influencing the Water Quality Index (WQI) in the region.

In conclusion, this research contributes to the evolving field of environmental science by delving into the temporal intricacies of water quality through the lens of time series forecasting. The anticipated outcomes promise to enrich our comprehension of water quality trends, providing a foundation for sustainable water resource management amid the burgeoning challenges of urbanization.

II. Literature Review

Traditional methods of water quality estimation have proven to be essential but come with inherent challenges, including high costs and time-consuming procedures. In light of these limitations, researchers have been exploring alternative methodologies that offer quicker and more cost-effective solutions.

Based on LSTM NN, Yuanyuan Wang et al.[1] developed an accurate prediction of a water quality indicator.The datasets were gathered by the Chinese Academy of Sciences' Nanjing Institute of Geography.Classifier models like as LSTM NN, BP NN, and OS-ELM have been used to compare outcomes from the same datasets.Compared to OS-ELM and BP NN, LSTM NN consistently has a greater prediction accuracy.The RMSE of BP increases gradually as the time step increases, however the LSTM NN has been very steady, staying lowest with each time step.In this paper, several simulations were run and parameter selections were made.With LSTM NN, more generalization is possible.But a lengthy training cycle necessitates the inclusion of another useful memory block.

Sang-Soo Baek et al.[2] used CNN to imitate water level and LSTM to analyze water quality.The Water Resources Management Information System in South Korea provided the data regarding water levels (WAMIS).Both CNN and LSTM used the mean square error (MSE) as the loss function for training their models.

A more efficient and straightforward approach to computing the water quality index (WQI) has been suggested by Umair Ahmed et al. [3].The program takes four inputs: total dissolved solids, pH, turbidity, and temperature.A wide range of supervised machine learning techniques are reviewed in this study.PCWR was the dataset's original source.

Q-value normalization is used to make index calculation easier.Z-Score normalization is used to convert all the data with various scales to the default scale.The dataset is adjusted to the previously specified scale.A few preparatory steps were taken before supplying the machine learning algorithm with the input.such as data partitioning and correlation analysis.In total, about fifteen different machine learning methods were applied.In the results analysis, polynomial regression was determined to be the most efficient technique with an MAE of 2.7273, MSE of 12.7307, and RMSE of 3.5680.The tested algorithms are unable to estimate the water quality based on real-time data.The process is shown in the figure below.

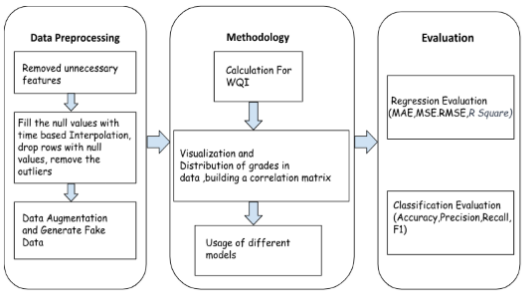

Fig. 1: Methodology Flow

III. DATASET

A. Source of Data

The dataset used in this investigation was gathered from the Bangladesh Water Development Board (BWDB). It spans a substantial time frame of 20 years, providing a comprehensive overview of water quality dynamics. The data comprises approximately 1400 rows, with all observations collected from monitoring stations strategically positioned along the river beside Dhaka.12 parameters are enlisted in Table 1.

TABLE I: parameters and their corresponding "WHO" standard limits

Parameter	WHO Limits
Alkalinity	500mg/L
Appeareance	Clear
Calcium	200mg/L
Chlorides	200mg/L
Conductance	2000 S/cm
FecalColiforms	NilColonies/100mL
Hardnessas CaCO3	500mg/L
Nitrite as Nitrogen Dioxide	¡1mg/L
pH	6.5-8.5
Temperature	Celsius
Total Dissolved Solids	100mg/L
Turbidity	5 NTU

B. Features Included

A variety of water quality metrics are included in the dataset, such as the date, temperature (°C), pH, dissolved oxygen (DO) in mg/L, salinity in ppt, and total dissolved solids (TDS) in mg/L.

C. Data Preprocessing

To guarantee the accuracy and consistency of the data, a comprehensive preprocessing process was applied to the dataset. In the beginning, NaN values were filled in with the median of each corresponding column in order to handle missing values. Then, using box plots and the interquartile range (IQR) approach, outliers were found and eliminated. To remove any potential distortions brought on by extreme values, this step was essential.It can be useful to test time series analysis techniques or look at how algorithms react over long periods of time by creating fake data.

Following data cleaning, the Water Quality Index (WQI) was calculated using a predetermined formula that included important water quality characteristics. Based on the determined WQI values, a matching grade was given to each data point, classifying the water quality into groups of one to five. This classification made the water quality easier to understand and helped with result interpretation..

By improving the dataset's integrity, these preprocessing procedures hoped to provide a solid basis for any further analyses and interpretations.We determined the WQI of each sample using equation (1), where, as shown in Table 2[4,5], wfactor represents the weight of a specific parameter and q is the value of a parameter in the range of 0–100.

$$\mathrm{WQI} = \frac{\sum \mathrm{qvalue} \times \mathrm{wfactor}}{\sum \mathrm{wfactor}} \tag{1}$$

TABLE II: Parameters weights for the WQI calculation

Weighting Factor	Weight
pH	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Solids	0.07
Nitrates	0.10
FecalColiform	0.16

D. Water Quality Class

After forecasting the WQI, we used the WQI in classification to define the water quality class (WQC) of each sample[4,5] as shown in Table.3

TABLE III: Ranges[4,5]

Water Quality Index Range	Class
0-25	Very bad
25-50	Bad
50-70	Medium
70-90	Good
90-100	Excellent

E. WQI Calculation

All data points were captured from monitoring stations strategically located along the river beside

TABLE IV: Descriptive Statistics for Selecting Water Quality Parameters

Parameter	Mean	Standard Deviation	Min	Max
Temperature (°C)	27.46	2.56	16.43	36.00
pH	7.30	0.63	4.26	9.50
Dissolved Oxygen (DO) (Mg/L)	4.52	3.49	0.00	23.10
Salinity (ppt)	0.19	0.21	0.00	1.70
Total Dissolved Solids (TDS) (Mg/L)	201.23	177.79	0.06	851.00
Electrical Conductivity (EC) (MicroS/cm)	458.01	320.47	4.06	1702.00
Turbidity	25.99	14.94	3.00	360.00
Iron (Fe) (Mg/L)	0.26	0.37	0.31	3.36
Chloride (Cl) (Mg/L)	49.20	37.53	0.00	198.00

Dhaka, ensuring a focused spatial coverage relevant to the study area. For WQI calculation,we need to find the pH Q,temp Q and tds Q values and DO Q values.After getting all these values we used the equation(1) to measure the WQI.Lastly,we assigned class to each WQI.Total number of class is 5.

1) Visulization and Distribution of Grades: To visualize the distribution of water quality grades of our dataset a heatmap is needed.The heatmap conveys an overview of how the different water quality grades in the dataset are spread out. Understanding the distribution of grades and spotting trends or anomalies in the data can be helpful.Examining the distribution of grades can provide information about the quality of water samples. It might show, for instance, whether the majority of samples fall into a particular grade range or whether there are variances.

2) Correlation Matrix: To determine whether there may be a correlation between the parameters, correlation analysis is done.Most importantly, a correlation matrix is needed to predict variables that are hard to measure and identify the dependent variables.Following the listing of the correlation analysis observations, we find that our prediction parameter WQI is related to four parameters: temperature, pH, total dissolved solids, and DO. TDS has a weak correlation with pH and temperature and a strong correlation with Fe and Cl.In order to reduce the system's cost, we must choose the bare minimum of parameters to measure the WQI.The four parameters that naturally choose themselves

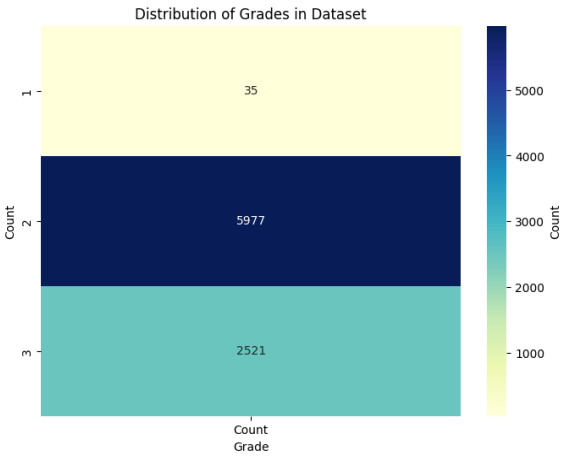

Fig. 2: Distribution of Data

are temperature, TDS, DO, and pH since they are the most affordable, easiest to monitor, and have the most effects on WQI. Total dissolved solids is another useful measure, and it is also easily accessible with a sensor.To summarize the correlation analysis, we used temperature, DO, pH, and total dissolved solids as the four factors for the WQI prediction. Table 4 provides a descriptive statistical analysis of the water quality parameter.

IV. Machine Learning Algorithms

A. LSTM

Long Short-Term Memory (LSTM) is an architecture for recurrent neural networks (RNNs) that

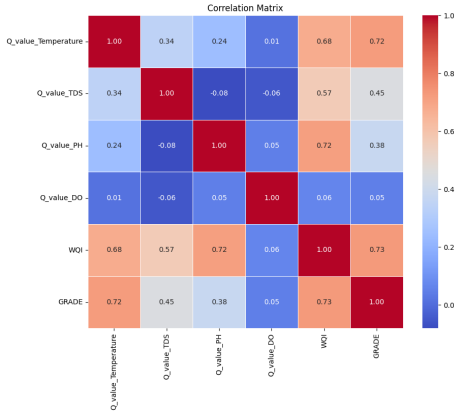

Fig. 3: Correlation Matrix

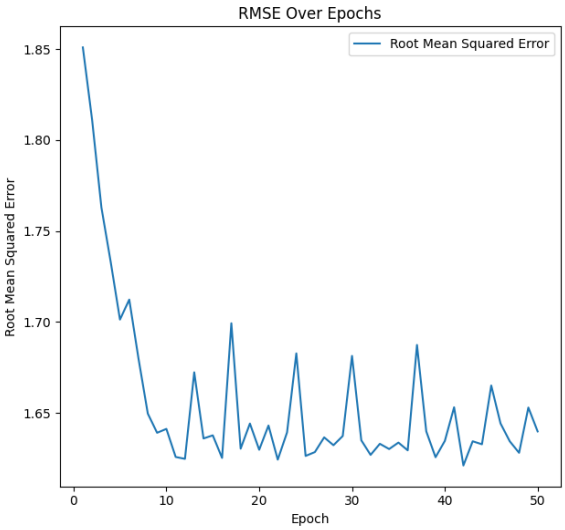

Fig. 4: RMSE over Epochs of LSTM

was created to address some of the shortcomings of traditional RNNs in terms of recognizing and comprehending long-term dependencies in sequential input.In order to implement LSTM, we preprocess the data and then applied stratification.By stratifying entire dataset is divided into mutually exclusive groups.It indicates that data should represent heterogenous population.Our dataset is divided into three subgroups train,test and validation.Then we applied the LSTM algorithm.For activation function ReLu is used.LSTMs successfully address the vanishing gradient issue and this enables LSTM's to capture and remember long-term dependencies in sequential data.

B. GRU

To improve the modeling of long-term relationships in sequential data and address the Vanishing Gradient Issue, a type of recurrent neural network (RNN) architecture known as a Gated Recurrent Unit (GRU) was developed.GRU simplifies the architecture compared to LSTM while maintaining comparable performance.GRU consist hidden state,update gate,research gate.The way we prepare the data for LSTM classifier in the same way GRU's data has been processed. However,GRU computationally less expensive than LSTM.We have used the built in GRU algorithm.

C. Result Analysis

Four factors were initially used: temperature, DO, pH, and total dissolved solids.This investigation carefully contrasted Long Short-Term Memory

Forecasting water quality indices in the surrounding rivers of Dhaka City is done through the use of Gated Recurrent Unit (GRU) and Long Short-Term Memory models. The results clearly support LSTM, which performs better across metrics. Lower mean absolute error (MAE) of 0.0170, mean squared error (MSE) of 0.0018, and root mean square error (RMSE) of 0.0421 for LSTM demonstrate its capacity to capture temporal nuances, making it the model of choice for accurate water quality predictions.

Despite yielding less favorable findings, as seen by increased MAE (0.8269), MSE (2.6660), and RMSE (1.6328), GRU nevertheless offers valuable information that should be taken into account in some situations.

Algorithm	MAE	MSE	RMSE
LSTM	0.0170	0.0018	0.0421
GRU	0.8269	2.6660	1.6328

V. Conclusion

In conclusion, our study aimed to enhance water quality index (WQI) forecasting in Dhaka City's peripheral rivers using advanced deep learning models—Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The results demonstrate the superior performance of LSTM in accurately predicting WQI values. Its capability to capture long-term dependencies and intricate temporal patterns positions it as the preferred model.

The robust performance of LSTM contributes significantly to water quality management, providing a powerful tool for reliable forecasts. While both LSTM and GRU showed efficacy, LSTM's distinct advantages make it the model of choice for nuanced temporal dynamics.

Future research may involve fine-tuning LSTM architectures and exploring additional features. In summary, the conclusive preference for LSTM underscores the importance of leveraging advanced deep learning techniques for accurate time series forecasting in environmental monitoring, contributing to sustainable water management.

References

[1] Y. Wang, J. Zhou, K. Chen, Y. Wang and L. Liu, "Water quality prediction method based on LSTM neural network", in 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 1-5, Nov. 2017

[2] Baek, S.-S.; Pyo, J.; Chun, J.A. Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach. Water 2020, 12, 3399. https://doi.org/10.3390/w12123399

[3] Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R. Efficient Water Quality Prediction Using Supervised. Water 2019, 11, 2210.

[4] Thukral, A.; Bhardwaj, R.; Kaur, R. Water quality indices. Sat 2005, 1, 99.

[5] Srivastava, G.; Kumar, P. Water quality index with missing parameters. Int. J. Res. Eng. Technol. 2013, 2,609-614.

[6] Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; pp. 1015–1021.

[7] Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359