Water Quality Prediction Using LSTM and GRU Models in Deep Learning

Chetan Likhitkar, Yash Kohinkar, Aniket Patil, Shubham Pawar

Chetan Likhitkar, D.Y. Patil University
Yash Kohinkar, D.Y. Patil University
Aniket Patil, D.Y. Patil University
Shubham Pawar, D.Y. Patil University
Professor Madhavi Patil, Dept. of Information Technology Engineering, D.Y. Patil University, Maharashtra, India

---***---

Abstract – *Water quality management is an important aspect of environmental sustainability that affects ecosystems, public health, and community well-being. Conventional water quality prediction methods often struggle to capture the dynamic temporal patterns inherent in environmental data. This project uses advanced deep learning models, especially Long Short-Term Memory (LSTM) and Gated Recurrent Segmentation (GRU), to solve time prediction problems in water quality monitoring.*

The project began by collecting comprehensive water quality data from a variety of sources, from sensors to satellite imagery. These data, including parameters such as pH levels, temperature, and dissolved oxygen, form the basis for the training and validation of LSTM and GRU models. Data processing techniques used to handle missing values, scale normalization, and build temporal series are necessary for effective deep learning.

The LSTM and GRU models were chosen for their ability to capture long-term dependencies, which are important for understanding the changing nature of water quality parameters. The architecture of the model is carefully designed, considering the input layer that accounts for temporal aspects, the hidden layer that captures complex patterns, and the output layer that produces predictions for water quality parameters.

Key Words: **Water Quality, Deep Learning, Long Short-Term Memory (LSTM), Gated recurrent Unit (GRU), Prediction model**

1.INTRODUCTION

This Water exceptional evaluation is a vital thing of environmental tracking, affecting public fitness, atmosphere stability, and aid management. Conventional water quality prediction techniques frequently depend upon empirical fashions and statistical evaluation, which may additionally have boundaries in capturing the complicated temporal dependence and nonlinear relationships inherent in water exceptional facts. In recent years, deep gaining knowledge of strategies have shown promising capacity in solving such complexities, offering a new paradigm for accurate and green water nice prediction.

Long quick time period memory (LSTM) and Gated Recurrent Unit (GRU) are excessive-stage recurrent neural community (RNN) architectures for modelling continuous records and handling long-term dependencies. These models have proven extensive fulfillment in numerous periodic forecasting troubles, making them suitable applicants for solving emerging water fine records challenges.

This paper offers a comprehensive analysis of the use of LSTM and GRU models for water high-quality prediction. Through leveraging deep mastering capabilities, our research objectives to enhance the accuracy and reliability of predictive fashions, thereby contributing to more powerful water quality tracking and management. The following section describes the database used, the method used, and the experimental results acquired, and affords information on the capability of LSTM and GRU fashions to develop water fine prediction methodologies.

2. Problem Statement

Water quality prediction is a critical aspect of environmental monitoring, essential for ensuring the safety and sustainability of water resources. Conventional methods for predicting water quality often face challenges in capturing the complex temporal dependencies and non-linear patterns inherent in water quality data. To address these challenges and enhance the accuracy of water quality predictions, there is a need to leverage advanced machine learning techniques, particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures in deep learning.

Identify the challenges in accurately predicting water quality over time. Emphasize the need for advanced predictive models to address the dynamic nature of water quality parameters.

3. Literature Survey

In the field of deep learning for water quality prediction, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models have garnered significant attention for their effectiveness in capturing temporal dependencies. Various studies have investigated the application of LSTM and GRU models across different domains, demonstrating their superior performance in prediction tasks.

Barzegar et al. (2020) proposed a hybrid CNN-LSTM model for short-term water quality variable prediction, illustrating the efficacy of deep learning in this context. Similarly, Zhang et al. (2019) conducted a study comparing LSTM, GRU, and standard RNN for water saturation prediction, highlighting the competitive performance of LSTM and GRU models. Additionally, Pan et al. (2020) developed a water level prediction model based on GRU and CNN, showing that the CNN-GRU model outperformed traditional methods like ARIMA and LSTM in terms of accuracy.

Furthermore, Haq & Harigovindan (2022) emphasized the importance of GRU and LSTM in predicting water quality for smart aquaculture. In their study, CNN was used for feature extraction, while GRU and LSTM were employed for learning temporal dependencies, underscoring the significance of these models in water quality prediction tasks.

In conclusion, the existing literature highlights the increasing interest in utilizing LSTM and GRU models for water quality prediction. These models have proven effective in capturing intricate temporal patterns and enhancing prediction accuracy compared to conventional methods.

4. Proposed System

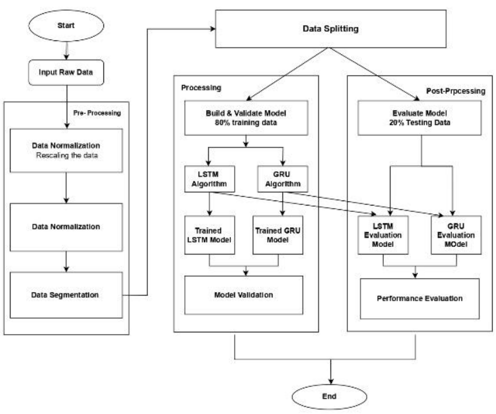


Fig :- Praposed Diagram

Fig -1: Proposed system diagram

The standards used to assess the sustainability of water resources are constantly evaluated as new factors are found. Standards and guidelines for contamination levels in drinking water are being developed by regulatory agencies. In response to the changing criteria, the water supply sector is creating new and improved operating and treatment procedures. All elements that affect water quality, as well as the public health relevance of components and available treatment technology, must be considered when developing drinking water quality guidelines.

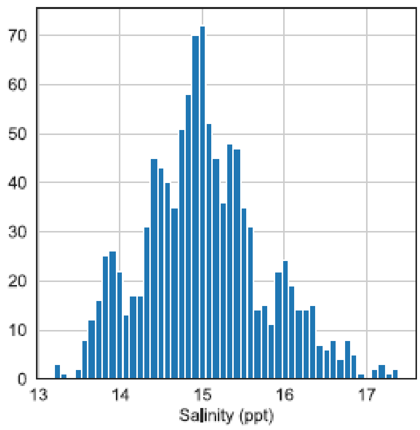
The initial task was to find out which factor would give a good indication of the quality of the water. Water parameters delve into the logic behind these choices. These measurements provide very little information about how dirty the water is on its own. As a result, the study will take into account the collective behavior of the parameters to produce a legitimate output, which will determine if the water is potable or not.

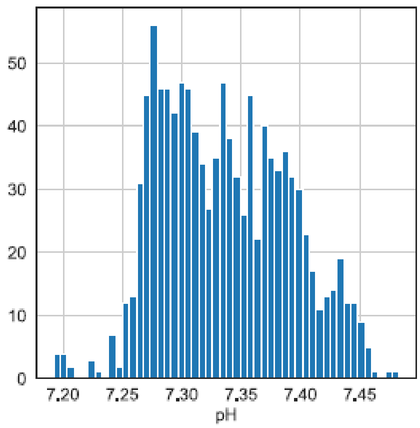
The second task was to deal with the dataset's missing values. The value of some factors may not be specified while defining the models, and the output may differ as a result. To solve this problem, we have included the mean value of the factor for which data is absent. To train the model efficiently, we first focus on data normalization using LSTM & GRU, which is a technique of data analysis. To achieve our goal, we appropriately calculate the Water Quality Index (WQI) to analyze water quality. For better representation, we provide a histogram of the dataset.

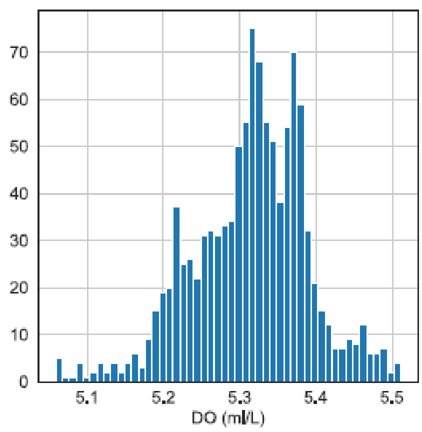
This facilitates for us to observe how our entire dataset is distributed. Then we have applied a correlation technique to determine the ability of two features to change at a constant rate. After that, we have split the entire dataset into two sections: training data and testing data. We used a variety of machine learning algorithms to train the dataset and then compare the models' accuracy. Following the application of those strategies, we employ hyper parameter tuning to evaluate and receive outcomes from our desired model. Finally, we use the accuracy of our suggested models to compare all of the results. As a result, the validity and reliability of our entire study are guaranteed by this approach. Figure 1 shows the flow diagram of the proposed model.

5. Helpful Hints

5.1 Figures and Tables







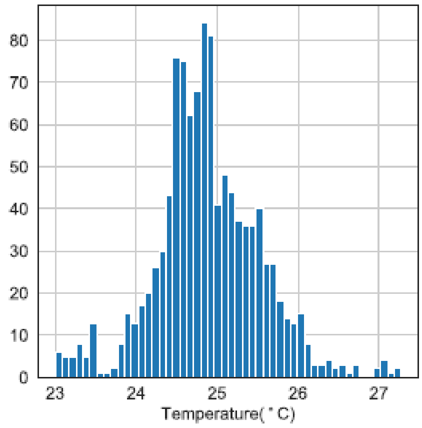


Fig -2: Water quality parameters variations trend of ADAK water quality dataset

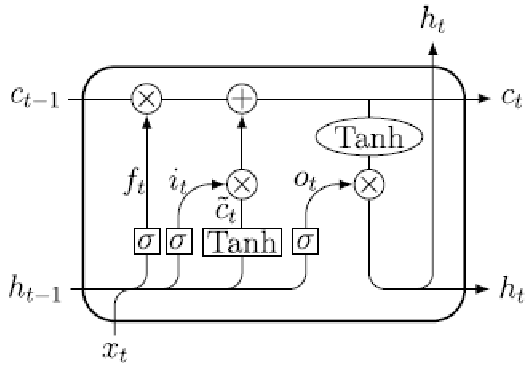


Fig -3: LSTM cell structure

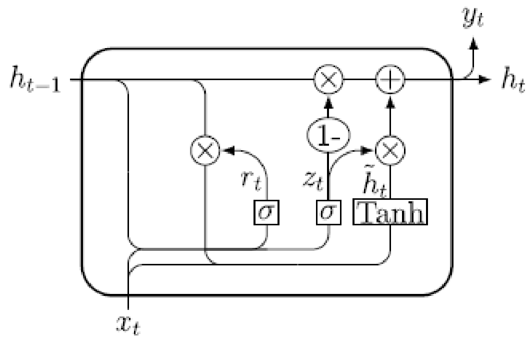


Fig -4: GRU cell structure

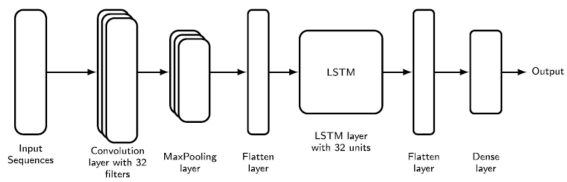


Fig -5: Proposed Hybrid CNN-LSTM DL neural network model structure

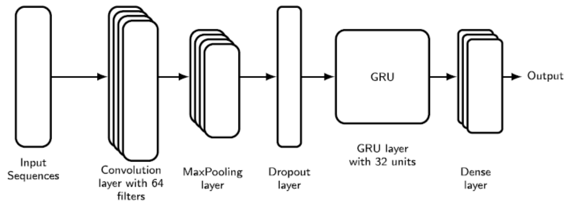


Fig -6: Proposed Hybrid CNN-GRU DL neural network model structure

Table -1: The Structure of proposed hybrid CNN-LSTM and CNN-GRU DL models.

Structure of CNN-LSTM and CNN-GRU DL models	
CNN-LSTM	ConvlD Layer (32 filters + 3 filter sizes, ReLU activation) Maxpooling (2 pooling size) Flatten () LSTM layer (32 units, ReLU activation) Flatten () Dense layer (1 neuron) Compile (MSE loss, Adam optimizer, lr = 0.0008)
CNN-GRU	ConvlD layer (64 filters + 5 filter sizes, ReLU activation) Maxpooling (4 pooling size) Dropout layer (0.2) GRU layer (32 units, ReLU activation) Dense layer (30 neuron, ReLU activation) Dense layer (10 neuron, ReLU activation) Dense layer (1 neuron) Compile (MSE loss, Adam optimizer, lr = 0.0008)

5.2 Experiments

5.2.1 Epoch

The result of choosing different periods {10, 50, 100, 150, 200, 300, 400, 500} were investigated for each DL model. Figure 6 shows RMSE vs time for various water quality parameters; they use the ADAK water quality database being proposed hybrid DL model and LSTM, GRU as well as CNN DL model. The hybrid DL model proposed here is CNN-LSTM and CNN-GRU maintains a better performance than the beginning DL models are LSTM, GRU and CNN. Hybrid model good performance and generation- Alization was achieved in 100 eons. Prediction accuracy performance is maintained for all four types of water-Period from 10 to 500. Also, prediction performance Our hybrid DL model is superior compared to our base model DL model. Like the base DL model at least 400 periods to achieve generalization and understanding the study data is less than 400 periods. Basic DL model is close to the performance of the proposed model 500 in the period. basic models require more computational resources to achieve similar results. Hybrid model For ADAK water, start a little after 150 cycles quality data set. However, accuracy is maintained, indicating- Another adaptation of the proposed DL model database.

5.2.2 Window Size

The effect of choosing a different window size {10, 20, 30, 40, 50, 60, 70, 80} for each DL model is analyzed. Giant. 8 shows RMSE vs window size for different waters quality parameters on the ADAK water quality data set s proposed hybrid DL models and basic DL models. The designed hybrid DL models perform better than the basic ones DL models. When the window size increases from 10 to 80, for hybrid DL models, the error remains constant. In this experiment, we see that these basic models are comparable in performance with hybrid DL models for some windows sizes for each water quality parameter. However, no baseline the models perform consistently well for all four waters qualitative parameters. For example, the performance of LSTM is slightly better than CNN-LSTM at window size 40 and 80 for pH. However, for the other three water quality parameters the performance of LSTM is not good. In comparison designed the hybrid models consistently perform well for the four parameters of water quality and window size that we have experimented with.

5.2.3 Learning Rate

Study level effect {0.01, 0.001, 0.0009, 0.0008, 0.0007, 0.0001} in performance for each DL model is analysed. In Figure 10, we compare RMSE vs. learn- performance level of the proposed hybrid DL model with the basic model using the ADAK water quality database. The proposed hybrid DL model consistently performs better basic DL model. However, the main indicator is performance DL model is not suitable for different training levels and various water quality parameters. Therefore, using the foundation of the DL model is not practical for WQP.

5.2.4 Batch Size

In this test, we analyse the performance of each DL different models and batch sizes {16, 32, 64, 128, 256, 512}. RMSE vs batch size for different watersheds quality parameters in the ADAK water quality database for hybrid DL model and basic DL model. The proposed hybrid can be compared to the DL model monitored the performance of the main DL model inconsistent and inferior. Also, we can see that CNN performance decreases with increasing party size.

5.2.5 Computational Time

In the four tests above, it's time to count measured and saved at the same time. Figure 14 and Figure 15 set the calculation time for each hyperparameter for ADAK water quality set and MAC water quality set, each other. Calculation time required by each DL model is different. Here we can observe some general trends from his place. One of them is calculation time increases with increasing period and window. But as the batch increases, the computation time decreases in size. We won't see much change in computing time for different study levels since we choose to study Estimated interval limits for two databases for all models.

5.2.6 Multi-Step Prediction

In this test, we analyze the performance of each {2,4,6,8,10,12} of the DL model for different step sizes. RMSE vs step size for different watersheds quality parameters in the ADAK water quality database for Hybrid DL model and basic DL models. The results show the proposed hybrid DL model performs better than the basic DL model. enhancements, moreover, the performance of CNN-LSTM and CNN-GRU decreases with increasing step size. RMSE vs step size for different Water quality parameters in the MAC water quality database with the proposed hybrid DL model and base DL model. From the results, better performance of the proposed hybrid CNN-LSTM model compared to other mainstream and hybrid models CNN-GRU model. Compared to the proposed hybrid DL model for two databases, the basic DL model not enough

5.3 Abbreviations

Abbreviations		
1.	WQP	Water Quality Prediction
2.	GRU	Gated Recurrent Unit
3.	LSTM	Long Short-Term Memory
4.	CNN	Convolutional Neural Network
5.	DL	Deep Learning
6.	RNN	Recurrent Neural Network
7.	ReLU	Rectified Linear Unit
8.	ConvlD	Convolutional Layer
9.	MaxpoolinglD	Pooling Layer
10.	IoT	Internet Of Things
11.	RMSE	Root Mean Square Error
12.	MAE	Mean Absolute Error

5.4 Equations

The LSTM architecture is shown in Fig. and memory units are defined as follows:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$$

$$\tilde{c}_t = \tan h\,(W_c \times [h_{t-1}, x_t] + b_c)$$

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tan h\,(c_t)$$

In this architecture f_t, i_t , and $o_t$are forget, input and output gate layer respectively. $\tilde{c}_t$ and c_t are new and final memory cell, w is weight matrices, b is bias vectors, $\sigma$is the sigmoid activation function.

GRU has fewer tensor operations than LSTM and runs typically faster than LSTM. The GRU architecture is shown in Fig. and memory units are defined as follows:

$$z_t = \sigma(W_z \times [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \times [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tan h\,(w \times [r_t \times h_{t-1}, x_t])$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t$$

In this architecture z_t is the reset and r_t is the update gate, $\tilde{h}_t$ is process input, and h_t is hidden state update, w is weight matrices and σ is the sigmoid activation function.

The performance of the prediction models is evaluated using mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE) and mean absolute percent-age error (MAPE), computed by the set of equations given below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(A_i - Y_i)|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(A_i - Y_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - Y_i)^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|(A_i - Y_i)|}{A_i}$$

Where A_t is the actual value of i_{th} sample, Y_i is the predicted value of i_{th} sample and n is the number of samples.

3. CONCLUSIONS

In this research work, we propose a DL model, CNN-LSTM and CNN-GRU for WQP. The prediction model was developed and tested in two separate databases. We also analyze the impact comprehensively horsepower conversion. Productivity ratio and beyond analysis HP is optimally used. We compare the performance of this means that DL models (CNN-LSTM and CNN-GRU) with basic DL models (LSTM, GRU and CNN) and focused based DL model (similarity-based LSTM and GRU based focus) MAE, MSE, RMSE, and card. The results of CNN- LSTM hybrid model shows a significant improvement in prediction accuracy as the computational time compared to the basic DL models. The hybrid model has similar performance compared to the focus-based model. However, the computational time is higher than the focus-based model and it offer a realistic solution for predicting water quality parameters.

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to the research community, colleagues, and organizations that have contributed to the development of the Water Quality Prediction Using LSTM and GRU models in Deep Learning presented in this research paper. This work would not have been possible without their dedication and expertise. Feel free to personalize and expand on this acknowledgment as per your specific situation and the contributions you'd like to recognize.

REFERENCES

[1] Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model Rahim Barzegar1, Mohammad Taghi Aalami2, Jan Adamowski3 2020Stoch Environ Res Risk Assess

[2] Potential for Prediction of Water Saturation Distribution in Reservoirs Utilizing Machine Learning Methods Qitao Zhang1, Chenji Wei2, Yuhe Wang3 et al. 2019Energies

[3] Water Level Prediction Model Based on GRU and CNN Mingyang Pan1, Hongzhan Zhou2, Cao Jun3 et al. 2020IEEE Access

[4] Water Quality Prediction for Smart Aquaculture Using Hybrid Deep Learning Models K. P. Rasheed Abdul Haq1, V. P. Harigovindan2 2022IEEE Access

[5] PCRWR. Water Quality of Filtration Plants, Monitoring Report; PCRWR: Islamabad, Pakistan, 2010. Available online: http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS%20REPOT-CDA.pdf (accessed on 23 August 2019).

[6] Ahmad, Z.; Rahim, N.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. 2017

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

[8] Amral, N.; Ozveren, C.; King, D. Short term load forecasting using multiple linear regression. In Proceedings of the 2007 42 nd International Universities Power Engineering Conference, Brighton, UK, 4–6 September 2007

[9] LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015

[10] Zhao, Y.; Li, J.; Yu, L. A deep learning ensemble approach for crude oil price forecasting. Energy Econ. 2017

[11] Le, X.H.; Ho, H.V.; Lee, G.; Jung, S. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* **2019**

[12] Gao, S.; Huang, Y.; Zhang, S.; Han, J.; Wang, G.; Zhang, M.; Lin, Q. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *J. Hydrol.* **2020**

[13] W. Zheng, P. Zhao, K. Huang, and G. Chen, ``Understanding the property of long term memory for the LSTM with attention mechanism,'' in Proc.30th ACM Int. Conf. Inf. Knowl. Manage., Oct. 2021