GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

Water Quality Prediction Using Artificial Intelligence Techniques

Dikshya Naik, CH Navina Prusty, Sneha Biju, Computer Science and Engineering, GIET University

Dr. Bidush Kumar Sahoo, Associate Professor, Computer Science and Engineering, GIET University

Abstract: Accurate prediction of water quality is essential for preserving freshwater resources and supporting effective environment management. In this study deep learning models, including Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) and Hybrid model of LSTM and GRU were applied for predict the water quality index(WQI). In addition, Extreme Gradient Boosting (XGBoost) and Random Forest (RF) type of Machine Learning Models are used for classify water quality class (WQC). The model was evaluating using standard statistical measures. GRU model delivered better prediction compare to the performance of LSTM and Hybrid approach, as indicated by higher R^2 value. While in WQC XGBoost achieve higher accuracy. The outcomes demonstrate the potential of combining Deep Learning and Machine Learning techniques to improve accuracy of water quality as well as management strategies.

Key Word – Water Quality Index, LSTM, GRU, Classification, Hybrid Model, XGBoost, Random Forest

1. Introduction:

As we know water is a fundamental necessity for ensuring of life, and its consumption is vital for almost all living organisms. However, ensuring the purity and safety for drinking water remains a significant challenge across the globe. Contaminated water is directly linked to numerous health issues such as waterborne diseases, gastrointestinal infections, and even long-term chronic conditions due to the presence of harmful substances. Moreover, poor water quality has broader environmental consequences, including the disruption of aquatic ecosystems and soil degradation. This makes monitoring, control, and management of water quality.

Traditionally, water quality has been assessed through manual testing methods, which rely on laboratory-based chemical and biological analyses. Parameters such as pH, Total Dissolved Solids (TDS), sodium, nitrates, Electrical Conductivity(EC), chlorides, and potassium are measured and evaluated to determine the safety of water for consumption. While effective, these manual approaches are time-consuming, costly, and often require skilled personnel. Additionally, such methods are not feasible for continuous real-time monitoring, leading to delays in decision-making. In many cases, authorities are able to identify contamination only after the damage has already occurred. Human dependency on manual interpretation of water parameters can also introduce inconsistencies and subjectivity, making it difficult to achieve a standardized and scalable solution.

The major challenge lies in the fact that contaminants like TDS, nitrates and chlorides are difficult to filter or remove completely once they reach high concentrations. Therefore, the need for predictive modelling of water quality becomes essential. Although existing studies

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

have attempted to calculate the Water Quality Index (WQI) using statistical and conventional ML methods, often fail to capture the temporal dependencies and nonlinear relationship among water quality parameters. Further, most models are limited to predicting WQI values but do not provide actionable classification into categories such as excellent, good, poor or unsuitable for drinking. This limits their practical utility for policy makers, environmental agencies and communities that require clear, real time decision making support.

To address these challenges, our work proposes the use of deep learning techniques especially Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and a hybrid model of LSTM and GRU for Water Quality Index Prediction. These models are particularly effective in capturing short as well as long term dependencies within time-series data, making them suitable for water quality analysis. By employing feature engineering and model tuning, we enhance the accuracy of prediction while minimizing computational complexity. Beyond regression-based prediction, this work also focuses on classification of water quality in two categories i.e. excellent, good, poor, very poor yet drinkable and unsuitable for drinking using advanced classifier like XGBoost and Random Forest.

This integrated approach not only bridges the gap between predictive modelling and practical categorization but also provide an effective tool for real time water quality monitoring. The outcome of this work can serve as a valuable decision-support system for individual community and governing authorities to ensure safe drinking water and safeguard public health.

1.1 Objective of the study

1. To develop and implement models of Deep learning (LSTM, GRU and Hybrid LSTM+GRU) for better prediction of Water Quality Index (WQI).

2. To classify the water quality into categories i.e. Excellent, Good, Poor, very poor yet drinkable and unsuitable for drinking, using algorithms of ML i.e. XGBoost and Random Forest.

3. To Performs data preprocessing, feature selection, and correlation analysis for identifying the most significant water quality parameters influencing WQI.

4. To provide a reliable and effective framework for real time water quality monitoring and decision-making support for safe drinking water management.

1.2 Scope of the study

This study focuses on developing a data-driven framework for predicting and classifying for the Water Quality Index (WQI) using selected physicochemical parameters. The scope includes data preprocessing, feature selection and the application of deep learning and machine learning techniques for regression and classification tasks. The work specially considers twenty-one parameters from the dataset, out of which seven highly correlated parameters are selected to improve efficiency and reduce computational complexity.

The study is limited to predicting WQI and categorizing water into different quality levels i.e. Excellent, Good, Poor, Very Poor Yet Drinkable and unsuitable for drinking. The primary emphasis is on the integration of prediction and classification to provide interpretable, real-

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

time insights that can aid communities, policymakers, and authorities in water quality management.

2. Literature Review

Artificial Intelligence(AI) techniques have increasingly been adopted for modelling and predicting water quality, owning to their capacity to handle complex, nonlinear and uncertain environmental data more efficiently than traditional statistical methods. Several studies have demonstrated the potential of hybrid and deep learning models for both forecasting Water Quality Index (WQI) and classifying water quality conditions.

For instance, a hybrid approach proposed for Prediction of Water Quality Using AI (MDPI, 2021) combined Adaptive Neuro Fuzzy Interference system (ANFIS) in regression with FeedForward Neural Network (FFNN) and K-Nearest Neighbour (KNN) for classification. The study highlighted that ANFIS outperformed other models in prediction accuracy, primarily due to its ability to manage fuzzy uncertainties and adaptively learn from data. Meanwhile, FFNN and KNN were effective in categorizing water quality into discrete classes, suggesting that integrating soft computing techniques can enhance overall water quality management strategies.

Similarly, prediction of quality of water using AI (ResearchGate, 2020) compared several AI-based models for WQI prediction. In this work, the Nonlinear Autoregressive Neural Network (NARNET) slightly supper passed Long Short Term Memory (LSTM) Networks in forecasting, as indicated by lower Mean Squared Error (MSE). For classification task Support Vector Machine (SVM) demonstrated the best performance with 97.01% accuracy.

Advancement in IoT have further expanded the scope of AI in environmental monitoring. Analysis Quality of water using LSTM type of deep neural networks in IoT environment (MDPI, 2019) presented an IoT-enabled LSTM model for real-time water quality prediction using parameters such as pH, TDS, temperature, and turbidity. The LSTM framework clearly find long-term dependencies in time-series data, enabling continuous and scalable monitoring solutions. This integration of deep learning with IoT infrastructure underscored the feasibility of smart water management systems.

In summary, the literature reveals a clear transition from conventional machine learning techniques (SVM, KNN, FFNN) toward advanced deep learning approaches (LSTM, Bi-LSTM, hybrid models) and their integration with IoT systems. While models like ANFIS and NARNET demonstrate strong regression capabilities, deep learning frameworks such as LSTM and Bi-LSTM offer superior handling of temporal dependencies, which is crucial for water quality time-series forecasting. However, challenges such as handling missing data, ensuring model generalizability across different water bodies, and optimizing real-time integration still remain. These gaps highlight the need for further research into robust, scalable, and interpretable AI-based solutions for water quality prediction and management.

3. Methodology

3.1 Dataset

We have conducted an analysis on water quality, here the dataset was selected from the Central Pollution Control Board (CPCB). This includes water quality monitoring data across

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

various locations in India with chemical and physical parameters measured over different years (2019, 2020, 2021, 2022). The dataset contains columns: Well_ID, State, District, Block, Village for Geographical identifiers of water sampling locations, Lattitude, Longitude for spatial coordinates for precise mapping, Water Quality Indicators like pH, Electrical Conductivity (EC), Bicarbonates (HCO_3), Sulfates(SO_4), Nitrates(NO_3), Total Dissolved Solids (TDS), Calcium (Ca), Carbonates ($C0_3$), Chlorides (CL), Magnesium (Mg), Sodium (Na), Potassium (K), Fluoride (F), Total Hardness (TH).

3.2 Data Preprocessing

For processing of the data, the dataset was first examined for null values. The outliers were then examined and removed by using the interquartile ranges, because of which certain feature like CO_3 was removed from the further process. One of the target i.e. 'Water Quality Classification' was then encoded into numeric format using Label Encoding. The features were then scaled by using MinMax scaler and a correlation analysis was done to determine the influence of the parameters on the target variables. Due to this process, out of 21 parameters in the dataset only 7 parameters showed high correlation with WQI, namely electrical conductivity (EC), bicarbonate (HCO_3), total hardness (TH), calcium (Ca), magnesium (Mg), sodium (Na), total dissolved solids (TDS).

3.3 Prediction of Water Quality Index

RNN models, namely Gated Recurrent Unit (GRU), long short-term memory (LSTM) and their hybrid model used for the prediction of Water Quality Index.

a. Long Short-Term Memory(LSTM) model

The Long Short Term Memory (LSTM) network was performed in this study due to its effectiveness in modelling of sequential and time series data with long dependencies. Unlike traditional recurrent neural networks (RNNs), which are prone to vanish and explode gradient problems, LSTM utilizes memory cells which is regulated by input gate, forget gate and output gates. As the model is capable of retraining relevant features over extended sequences while discarding repeated one, it is selected as a suitable architecture as shown in figure 1. Configuration details of different layer is summarized in Table 1.

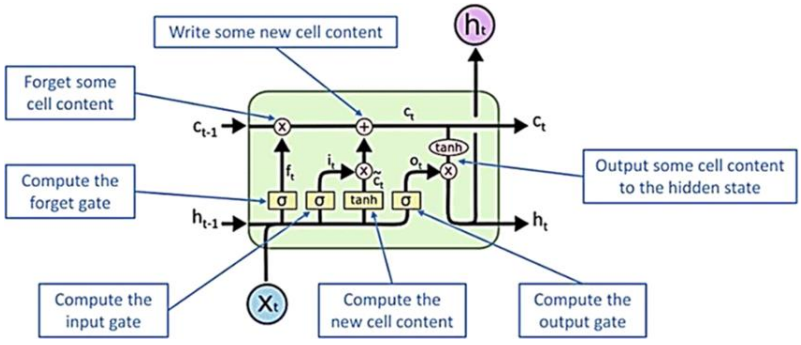

Figure 1: Cell Architecture

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

Table 3.1: Configuration details of different layer

Layers/Component	Configuration Details
LSTM Layers	64 units, tanh activation
Dropout Layers	20% dropout rate
Dense Output Layer	1 unit with sigmoid activation function
Loss Function	Binary Cross Entropy
Optimizer	AdamW
Training Configuration	Trained by 100 epochs of batch size of 32 and 20% validation split

b. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) is a type of RNN model, which is designed to capture longterm dependencies in sequential type of data efficiently. It is an extension of traditional RNNs and shares similarities with LSTM network. GRU model as shown in Figure 2, consists of two gates that is update gate for calculate using the entered data at time-step and the previous hidden state and second Research Gate for calculate in a same manner to the update gate. The architecture of the GRU model is shown in Table 2.

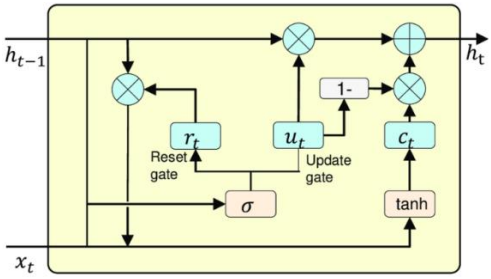

Figure 2: GRU Model

Table 2: GRU architecture

Layers/ Components	Configuration Details
GRU Layer	64 units, tanh activation
Dropout Layer	20% dropout rate
Dense Output Layer	1 unit with sigmoid activation function
Loss Function	Binary crossentropy
Optimizer	Adam
Training Configuration	Trained by 100 epochs with batch size of 32 , and 20% validation split

c. Hybrid Model (LSTM+GRU)

Some tasks might have diverse data dynamics certain parts of the sequence might benefit from LSTM's robustness in longer dependencies, while others might need GRU'S faster responsiveness. A hybrid model can adapt to these varying needs. The Architecture of Hybrid model is explained in the Table 3.

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

Table 3: Architecture of Hybrid model

Layers	Details
LSTM Layer	64 units, tanh activation, L2 regularization(0.001), return sequences, receives input of shape(timesteps, features)
Dropout Layer	30% dropout rate
GRU Layer	32 units,tanh activation, L2 regularization(0.001), return sequences
Dropout Layer	30% dropout rate
LSTM Layer	16 units, tanh activation, L2 regularization(0.001), does not return sequences
Dropout Layer	30% dropout rate
Dense Layer	8 units, ReLU activation
Dense Output Layer	1 unit, linear activation(for regression output)
Training Configuration	Optimizer: Adam, Epochs : 100, Batch Size : 32 and Validation Split : 20%

4. Findings

The study of importance of developing reliable and efficient methods for assessing water quality, as traditional laboratory-based testing is often time-consuming, costly, and unsuitable for real-time monitoring. Through systematic analysis, it was found that only a limited set of seven highly correlated parameters of pH, Electrical Conductivity (EC), Total Dissolved Solid (TDS), Carbonates and Chlorides was sufficient to generate accurate predictions of WQI. This reduction from the original twenty-one parameters not only improved efficiency, making the approach more practical for real-world applications.

Another significant finding is that the framework developed in this work does not merely predict numerical water into meaningful categories such i.e. Excellent, Good, Poor, Very poor yet drinkable and Unsuitable for drinking. This step is crucial because it translates raw data into actionable insights that can directly guide communities, health authorities, and policy makers in decision-making. Unlike tradition methods that provide delayed results, this approach enables near real-time monitoring, making it more efficient in identifying contamination risks before they escalate.

The contribution of this study lie in offering a comprehensive, data-driven solution for water quality management. It combines prediction with classification to ensure both accuracy and interpretability, applies feature selection to simply models without compromising reliability and presents a scalable framework for continuous monitoring. Overall, the work provides a valuable decision-support tool that can enhance public health protection and environmental management by ensuring safe drinking water availability.

5. Results

The result of this study demonstrate the potential of combining regression and classification techniques for effective water quality prediction. In regression performance, the GRU model provided balance of accuracy and reliability with the lowest values of RMSE and MAE,

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

slightly outperforming LSTM while predicting the Water Quality Index as shown in Figure 3. Although hybrid model gave a marginally higher R^2 value, its large error rates indicated instability and poor generalization, highlighting that increase complexity does not always guarantee improved performance. This suggests that simpler recurrent models can be more effective when tuned appropriately for structured datasets contains water quality parameters as shown in the Table 4.

Table 4: Water quality parameters for different Model

Models	RMSE	MAE	R^2
LSTM	34.509797	27.119982	0.524402
GRU	33.280663	26.171416	0.549795
Hybrid(LSTM+GRU)	115.273096	109.906191	0.552052

In terms of classification, Both Random Forest and XGBoost achieving the highest overall accuracy (99.33%), making it highly suitable for categorizing water into Excellent, Good, Poor, very poor yet drinkable, unsuitable for drinking types of meaningful classes. These findings not only validate the capability of ML and DL in predicting and classifying WQI but also emphasize their practical relevance for real-time monitoring, where accuracy, interpretability, and reliability are critical for decision-making. The values are showed below in the Table 5.

Table 5: Comparison of Random Forest and XGBoost.

Models	Accuracy (%)	Precision (%)	F1-score (%)
Random Forest	97.85	97	98
XGBoost	99.33	98	97

Figure 3: Model performance comparison

GIET UNIVERSITY JOURNAL, 2025, VOL. 5, NO. 1, PP. 18-25
Manuscript received 10th January, 2025; revised 20th May, 2025; Accepted 22nd June, 2025

ISSN: 2583-5955

6. Conclusion

In this study Deep Learning Models including LSTM, GRU and Hybrid LSTM-GRU architecture, for prediction of WQI. In addition, i.e. XGBoost and Random Forest of ML algorithms were applied for the classification of WQI data. The performance of the proposed models was remarked using standard statistical measures. The result indicated the the GRU model provided superior predictive capability compared to other, reflected with higher R^2 value. In the classification method XGBoost provided greater accuracy while finding water quality.

References

1. Prasad, D. V. V., Venkataramana, L. Y., Kumar, P. S., Prasannamedha, G., Harshana, S., Srividya, S. J., & Indraganti, S. (2022). Analysis and prediction of water quality using deep learning and auto deep learning techniques. Science of the Total Environment, 821, 153311.
2. Aldhyani, T. H. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). [Retracted] water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics, 2020(1), 6659314.
3. Liu, P., Wang, J., Sangaiah, A. K., Xie, Y., & Yin, X. (2019). Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability, 11(7), 2058.3. Dhamija, P. (2012). "E-recruitment: A roadmap towards e-human resource management." Research Journal of Management Sciences, 1(1), 23–26.
4. Khullar, S., & Singh, N. (2022). Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. Environmental Science and Pollution Research, 29(9), 12875-12889.
5. Hussein, E. E., Jat Baloch, M. Y., Nigar, A., Abualkhair, H. F., Aldawood, F. K., & Tageldin, E. (2023). Machine learning algorithms for predicting the water quality index. Water, 15(20), 3540.
6. Wu, J., & Wang, Z. (2022). A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. Water, 14(4), 610.