International Research Journal of Education and Technology

Peer Reviewed Journal, **ISSN2581-7795**

Enhanced Surface Water Quality Prediction Through LSTM-Enabled Deep Learning Techniques

Regina M[1], Chrisma Sirumani A[2], Raphael A[3]

[1]Assistant Professor, Department of Computer Science, Loyola College, Chennai, Tamilnadu

[2]M. Sc. Computer Science, Department of Computer Science, Loyola College, Chennai, Tamilnadu

[3]M. Sc. Computer Science, Department of Computer Science, Loyola College, Chennai, Tamilnadu

Abstract

Accurate prediction of surface water quality plays a vital role in environmental monitoring and ensuring public health. Traditionally, manual sampling and statistical approaches have been employed for this purpose; however, these methods are time-intensive and often fail to capture intricate patterns present in water quality data. This study leverages an LSTM-based deep learning model to enhance the precision of the Water Quality Index (WQI) prediction. Surface water quality is affected by factors such as seasonal fluctuations, pollution incidents, and climatic variations. Unlike conventional models that primarily account for short-term relationships, LSTM effectively captures long-term dependencies through its gating mechanisms (forget, input, and output), enabling it to emphasize significant trends while minimizing noise. Missing values in the dataset are addressed using mean imputation, and data preprocessing is carried out with MinMaxScaler for feature normalization. The model's performance is evaluated using the R^2 score, achieving a high accuracy of 99.9%, demonstrating the effectiveness of this approach.

Keywords:LSTM; MinMax Scaler; R^2 Score;Surface Water; WQI.

1. INTRODUCTION

Surface water refers to water bodies like rivers, lakes, ponds, reservoirs, and streams. It provides useful drinking water, agriculture, industry, and recreational activities too, and is, therefore, fundamentally important. Surface water is a major factor that affects the availability and the quality of water required for the maintenance of the balance of the ecosystems of the area, species biodiversity, and human health too.

In contrast, surface water is facing much impact due to the consequences of human activities and environmental changes. This has made it extremely important that people take on more effective and sustainable practices that will be necessary in the management and conservation of the environment. Surface water, even though very important, is threatened by lots of different contaminations such as industrial waste, runoff from agriculture, and untreated sewage. Heavy metals, pesticides, and microbial contaminants lead to the degradation of water quality, which in turn is not safe for consumption but also damages aquatic life. Climate change and urbanization are resulting from these challenges too, in such a way that water availability can be decreased and the water quality can be worsened. A solution for these problems is sustainable water management that includes monitoring, pollution control measures, and the use of sustainable techniques.

Water is considered to be polluted if unwanted substances change it in such a way that it is no more natural and is harmful to the environment. Water bodies can be brought polluting chemicals, plastics, and biodegradable waste into the surface and groundwater systems from industrial effluents, agricultural activities, and city runoff. This pollution is the reason why the processes of life become impossible, and as a result, aquatic biodiversity decreases, and people's health may be damaged due to the consumption of this polluted water.

Water pollution is a cause of many environmental and associated economic and social challenges. Water sources that are contaminated carry the expenses of water treatment, thereby limiting the possibility of clean drinking water. The polluted water bodies which decrease the productivity of the industries of both fisheries and tourism are likely to cause them economic losses. In order to reduce these dangers, water quality standards, clean-up of

the sources of pollution, and teaching the public to look after water resources are absolutely necessary.

The Water Quality Index (WQI) is a tool designed to check the water quantity and differentiation in terms of positive and negative aspects in different contexts based on various physicochemical and biological parameters determined. WQI simplifies extremely long and detailed water data descriptions by using a numerical scale, which is easier to understand for policymakers, researchers, and environmental protection agencies to explain the health of water bodies.

WQI is mainly used to find out pollutant trends and determine the degree to which water is suitable for drinking, irrigation, and industrial use. For evaluating water quality, there are different types of WQI that use various criteria. The National Sanitation Foundation WQI is one of the primary measurements and includes conditions like pH, turbidity, and dissolved oxygen, while the Canadian Council of Ministers of Environment WQI (CCME WQI) provides a version that can be adapted depending on the intended use of the water.Several WQI exist, and they use different criteria to assess water quality. Governments and organizations use these indices to implement necessary interventions to maintain and improve water quality.

The technologies of Machine Learning (ML) and Artificial Intelligence (AI) play pivotal roles in tracking and forecasting water quality. Assessing water quality is extremely challenging, requiring the analysis of vast datasets to derive conclusions. By analysing large datasets, AI models can detect patterns, predict contamination risks, and provide early warnings about water quality deterioration. A common way to automate water quality monitoring is through regression models and classification methods like neural networks, which involve less reliance on traditional laboratory testing. AI-based models have outperformed traditional models in estimating water quality indices. Rana et al. [6] compared the application of AI for surface water quality analysis with traditional statistical measures and deep models, achieving significant advantages with deep models.

Water quality prediction models enhance decision-making. These models can combine data from several sources such as sensors, satellites, and historical data to predict future water quality accurately. Through the use of AI, environmental policymakers and managers can take proactive action on water pollution before it happens and aid in effective and sustainable water management.

An example of a model used for water quality prediction is LSTM. Multiple studies have demonstrated that deep learning, particularly Long Short-Term Memory (LSTM) networks, provides maximum accuracy in predicting water quality indices. LSTM can learn sequences of signals over time and operate effectively with incomplete or noisy data.

LSTM uses memory cells and gates to control information flow, solving recurrent neural network (RNN) issues. Gates determine what information gets through, such as the forget gate controlling what information is remembered. It combines these gates with memory cells that enable the storage of information. LSTM has high relevance over long sequences, making it ideal for water quality predictions, weather forecasting, and routine pattern identification.

Zhou et al. [2] discussed the performance of LSTM models in extracting long-term dependencies from time series data, confirming their superiority over machine learning models like Decision Trees (DT) and k-Nearest Neighbours (KNN). Ahmed et al. [3] explored hybrid AI models that integrate different machine learning techniques to enhance prediction accuracy, demonstrating that combining LSTM with other deep architectures improves classification and prediction performance.

Managing missing values and outliers is crucial in water quality prediction. Low-quality data can harm model performance, necessitating effective imputation techniques. Babu and Reddy [1] analyzed various imputation methods for time series forecasting and found that mean imputation was the most effective. To enhance water quality determination, Khan et al. [4] developed an AI-based smart water monitoring system that employed imputation techniques to address discrepancies in sensor data.

Choosing specific attributes is also essential to increasing the model's accuracy. The most crucial water quality metrics may be found, which helps to improve prediction models. In line with the results of this study, which indicated BOD to be the most significant predictor, Shams et al. [7] found that

Biochemical Oxygen Demand (BOD), pH, and Dissolved Oxygen (DO) were the most critical parameters for determining water quality index values. Ahmed et al. [5] further noted that removing irrelevant features improved model performance and reduced computational costs.

One of the most widely discussed topics in recent years is the comparison between machine learning and deep learning for water quality prediction. Studies have suggested that deep learning techniques, particularly LSTM networks, significantly enhance prediction outcomes compared to conventional machine learning approaches. Abbas et al. [8] reported that LSTM models achieved substantially higher accuracy than Decision Trees and Support Vector Regression (SVR).

The increasing significance of deep learning, especially LSTM networks, in improving water quality forecasting is evident from existing literature. Previous studies emphasize the importance of feature selection, confirm that LSTM models outperform traditional approaches, and demonstrate the effectiveness of imputation techniques in handling missing data. These findings provide the foundation for future advancements in AI-based water quality assessment and management systems.

2. METHOD

This research applied a deep learning approach to derive Surface Water Quality Index (WQI) estimations. The use of traditional statistical methods to estimate water quality parameters is often unproductive because of the lack of complex relationships that are often missed. To solve this problem, we employed a Long Short Term Memory (LSTM) model, which is designed to retain long-term relationships in sequential data.

2.1.Dataset

The drinking water quality assessment data used in this work was obtained from Kaggle and the Government of India collected the data between 2005 and 2014 in different lakes and rivers. A total of 1,991 instances are contained in the dataset while seven features namely DO, pH, Conductivity, BOD, Nitrate, Fecal Coliform, and Total Coliform are present. The listed features enable crucial water quality measurement.

2.2 Mean Imputation

Dealing with missing values is a substantial duty when the data is being pre-processed in order to keep the dataset as a whole so that it can be analysed properly. Imputation is the method used in the missing value analysis of this study, using imputed values instead of missing values. The application of a reliable imputation method depends on the dataset type and data integrity, respectively.

Variation of imputation techniques in data science is well established, among which Mean Imputation (restored missing values by using the mean value of the column), Median Imputation (replaced the mean with the median to counteract the skewness of the data that resulted in using a mean), Mode imputation (replaced missing data with the most frequently occurring value in the column), K-Nearest Neighbours (KNN) Imputation (imputed missing values by looking at the nearest neighbours in the data), and Regression Imputation (missing values estimated by regression models based on other features available) represent the missing data methods. For the present research, mean imputation was chosen out of all because it is simpler and more effective means of dealing with discontinuous data.

In the context of the presented data set, mean imputation is recognized as the most ideal technique for imputing the missing values. The approach was implemented because the data set contained continuous numerical items and calculating the mean was the most ideal estimator to use for replacing missing values without causing substantial fluctuations to the entire value distribution. The process took the mean point among the existing items for the same column in filling the missing gaps. This will keep the database statistics the same but in a more efficient and secure way. Along with this, this mostly copes with such issues as the non-existence of missing values and the absence of database bias.

This study used mean imputation because it is relatively simple and effective in cases where time series data from sensors is continuously lost. This way the maintenance of the overall data distribution without the loss of meaningful information due to missing entries is possible. Additionally, since the data set deals with water quality parameters, the mean is a way of replacing missing values with reliable and authentic data.

The method used to perform mean imputation was as follows: 1) Missing Value Identification: Searched the dataset for missing values in the features selected to be utilized in the computation of the Water Quality Index (WQI); 2) Calculate the Mean: Calculated the mean for all seven of the water quality parameters selected; 3) Replace Missing Values: Replaced missing values in every column with calculated mean of every column; 4) Verify the Data: Ensured that there were no remaining missing values and that the dataset was statistically consistent. This procedure ensured data completeness while ensuring that water quality could be analysed reliably.

2.3 Water Quality Index (WQI) Computation

Water Quality Index (WQI) is a numerical index that represents the overall quality of water by integrating various water quality parameters. The Weighted Arithmetic Water Quality Index (WAWQI) method is used to determine WQI in a way that very crucial parameters such as Dissolved Oxygen and Biochemical Oxygen Demand contribute more to the resulting WQI than less crucial parameters such as conductivity.

The following seven water quality parameters were used to calculate WQI: Dissolved Oxygen (D.O.) (mg/L), pH, Conductivity (µmhos/cm), Biochemical Oxygen Demand (B.O.D.) (mg/L), Nitrate (Nitrate-N + Nitrite-N) (mg/L), Fecal Coliform (MPN/100ml), and Total Coliform (MPN/100ml).

The quality rating scale (qi) for each water quality parameter (i) is computed using the formula:

$$qi = 100 \times ((vi - vid) / (si - vid))$$

Where,

vi = Measured value of the parameter

vid = Ideal value of the parameter for pure water

si = Standard permissible value for the parameter

The unit weight (wi) for each parameter is determined as:

$$wi = k / si$$

where:

k is the constant of proportionality, given by:

$$k = 1 / (\Sigma (si))$$

N is the number of water quality parameters.

The overall WQI is computed using the weighted sum formula:

$$WQI = (\Sigma (qi \times wi)) / (\Sigma (wi))$$

This results in a single numerical value that represents the overall water quality.

Though Temperature is present as a column in the dataset, it is not considered as a feature for Water Quality Index Prediction. It reduces the oxygen solubility in water, which affects Dissolved Oxygen (D.O.). It affects pH by affecting dissociation of carbonates and bicarbonates and Biological Oxygen Demand (B.O.D.) by affecting microbial activity and organic decomposition. Since temperature already affects some of the other parameters included under WQI, it is not considered separately in the equation.

The WQI models such as NSF-WQI and Weighted Arithmetic WQI assign more weightage to parameters having direct impact on water quality and human health. Temperature is not regarded as an independent parameter in these models. Natural variation of water temperature occurs due to geographical location (tropical or temperate zone), season, and day of the week. Since such natural variations may or may not indicate pollution, incorporation of temperature in the WQI calculation may lead to misinterpretation.

WQI is a consistent water quality index in which different parameters are combined into a single value. Weighted Arithmetic WQI method is applied for ensuring the most important parameters have an adequate contribution to the final index to facilitate effective water quality monitoring and management.

2.4 Outlier Removal

Outliers are data points that differ from the rest of the dataset. Outliers can arise due to data corruption, genuine variability in data or measurement errors. Removal of outliers improves the model's performance by preventing extreme values from misinterpreting predictions. It amplifies data integrity by making sure the dataset precisely represents underlying patterns, helps in better visualization by making graphs more interpretable.

Asymmetry of a dataset's distribution is measured by Skewness and high skewness represents the presence of extreme values, making the detection of outliers a crucial one. Left-skewed data has extreme values on the left, which affects the mean and variance, while right-skewed data contains high value outliers that may misinterpret the analysis.

High Skewness (>3) proposes severe asymmetry, making the removal of outlier methods essential. Performing analysis of skewness helps determine whether removal of outlier techniques like Interquartile Range (IQR) or log transformations are needed to normalize the distribution of data.

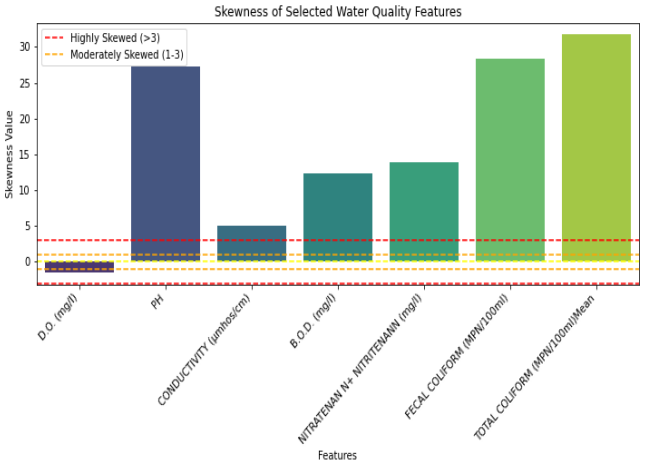

FIGURE 1. Skewness of Selected Water Quality Features

Analysis of skewness helps in understanding the data distribution and detecting outliers. From the dataset, the highly skewed features are pH (27.23), fecal coliform (28.28), total coliform (31.69), nitratenan + nitritenan (13.92), B.O.D (12.39), and conductivity (5.04). Severe asymmetry and presence of extreme values are indicated by these values. D.O. (mg/l) is a moderately skewed feature in the dataset with a skewness of -1.46 indicating slight left-skewness.

FIGURE 1 shows the amount and direction of skewness for each feature.Interquartile Range is chosen for outlier removal because most features are highly skewed, IQR works well for skewed data, robust against extreme values and it is simpler and more interpretable. The skewness values help in making the data distribution understandable. A skewness between 0 to ±0.5 indicated an approximate symmetric which is a normal-like distribution. A skewness between ±0.5 to ±1 propounds slight skewness, while values between ±1 to ±3 indicate moderate skewness. The data is highly skewed when the skewness is greater than ±3 which indicates severe asymmetry and the presence of extreme values.

Steps to remove outlier through Interquartile Range:

Step 1: Compute Quartiles:

Q1 (First Quartile): 25th percentile of the data has been taken as first quartile whereas the remaining 75th percentile of the data is taken as Q3 (Third Quartile).

Step 2: Interquartile Range (IQR) Calculation:

Interquartile Range (IQR) is calculated using this formula

$$IQR=Q3-Q1$$

Step 3: Determining Outlier Boundaries:

Two types of boundaries: Lower and Upper Bound.

$$Lower\ Bound=Q1-1.5*IQR$$

$$Upper\ Bound=Q3+1.5*IQR$$

Step 4: Identify and Remove Outliers:

An outlier is considered when any data point which is outside the lower or upper bound and it can be removed.

2.5 Data Normalization

Data normalization is one of the most important procedures within the machine learning area that changes the natural data into a common form in order to get better model performance. It means that different numerical features have similar ranges; thus, greater numbers cannot be so influential over the learning process. Specifically for deep learning models like LSTMs, normalization has always been necessary because different feature scales may induce convergence or instability. Through the process of normalizing data, the model gains proficiency in the learning union and it interpolates very well with fresh data.

Scaling is a really important step in the preprocessing stage of machine learning that changes feature values into a certain range, making them the same and maintaining it better. Thus, it helps in delivering numerical comparisons more easily and saves larger magnitude features from overshadowing the learning process. There are two principal scaling methods: Min-Max Scaling and Standard Scaling. Min-Max Scaling reshapes data within a specific range, such as 0 to 1 or -1 to 1, by scaling the values in that range accordingly. The approach works very well when the main objective is the preservation of the value relationships among the features and making all features contribute equally to the success of the model. Standard Scaling, known as Z-score normalization just like

Droving dibble, centres the data around zero and scaling it to have unit variance. This flexibility is best for data that commonly follows a normal distribution, so it guarantees that all features have the same scales and do not depend on abnormal numbers. The comparison of the two methods relies on the data set and the characteristics of the machine learning model.

In this study, MinMaxScaler was used for feature scaling. MinMaxScaler transforms features by scaling each feature to a given range. MinMaxScaler scales and translates each feature individually such that it is in the given range of the training set (between 0 and 1). The formula for MinMaxScaler is:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X is the original value, X_{min} is the minimum value in the column and X_{max} is the maximum value in the column. This modification assures that the data does not surpass the boundary, which in turn, makes it simpler for neural networks to process.

MinMaxScaler was preferred rather than the other procedures that are scaling because it protects the natural data distribution and, in the meantime, ensures that all feature values are within a definite range. This is essential for any type of deep learning model (e.g., LSTMs) since it minimizes the problem of large numerical differences that can affect and result in slower convergence speed of the gradient updates. If all input features are within the range of 0 to 1, the model will learn efficiently and as a result, the risk of both exploding and vanishing gradients diminish, enhancing the predictive performance of the model.

2.6 Data Splitting

Data splitting is a crucial phase of the implementation of machine learning that would lead the grounds of data which includes division of the dataset into some of the parted subsets for the training and the testing step to make sure about the process of the model's capability. In our study, we have taken advantage of the train_test_split function from scikit-learn to divide the dataset into 75% training data and 25% testing data. The part for training is used to give the model an ability for learning to make the right guess based on input features, while the part for testing is to know whether the model can generalize to a new question. In this way, it is also providing the advantage of improving the test result in order to make the accuracy suitable to decrease the inaccuracy, therefore, improving the effectiveness of the model. To confirm reproducibility, we have stayed at the random state of 0, which means that every time the code runs, the dataset is split in a certain way. Both the input features (x_train, x_test) and the target variable (y_train, y_test) were selected and divided before splitting to make sure that the structure of the subsets is the same. This kind of data splitting strategy is effective at the fact that the model is reviewed under unseen sample test data thus it offers to predict the performance as accurately as possible.

2.7 Long Short Term Memory (LSTM) Model

LSTMs, or Long Short-Term Memory, are artificial neural networks used for sequential data types, such as time series, speech, and text. It acts as an advanced version of an RNN and thus surpasses almost all the limitations of traditional RNNs input-level in handling long-term dependencies present in data. Three main building blocks, called gates, enable the LSTM to remember significant data and forget unimportant data. The Forget Gate decides what information from the previous time step should be discarded, the Input Gate decides any new information to be stored, and the Output Gate decides what information should proceed forward. These gates in an LSTM architecture allow it to hold a memory cell capable of keeping past information for longer, such as predicting stock prices, analysing text, or even forecasting weather trends.

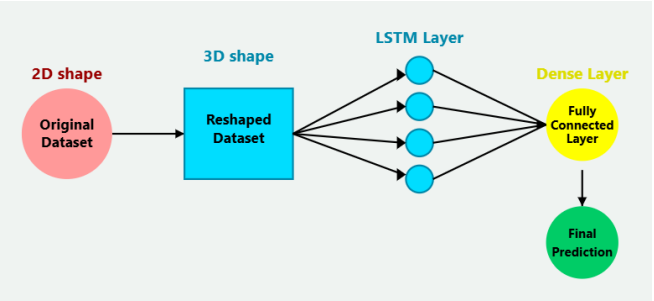

FIGURE 2.LSTM Model

FIGURE 2 shows the LSTM model workflow of sequential data processing. The dataset starts as a 2D thing (raw data) and that is handed to the model directly. LSTM networks have 3-dimensional input,

so the data is formatted for that much more. This transformation makes the time step dimension to be used in a proper way for pattern recognition in sequence. Next, this reshaped dataset is fed to the LSTM layer that has multiple LSTM units for learning the long-term dependencies in a data. Finally, the output from the LSTM Layer is piped to the final fully connected dense layer to identify the learned patterns better. Next the model outputs its final prediction which, may be for example, to be used in any task such as forecasting and classification or anomaly detection. With this type of structure, it makes the LSTM network work in a very efficient way with time-series and also keeps temporal relationships.

Table 1. Hyperparameter Tuning

PARAMETERS	REMARKS	VALUES TRIED
LSTM Units	Memory cells	100,150,200, 225,250
Batch Size	Batching	1,2,5,10
Epochs	Training cycle	200,300,350, 450,500
Optimizer	Weight Updates	Adam, RMSprop,Nadam, Adamax
Loss function	Error Minimization	MAE, MSE, Huber
Feature Scaling	Method for feature normalization	Standard Scaler, MinMax Scaler

Table 1 shows the different hyperparameters that were researched in the training of the LSTM model to find out the best configuration. A specified range was employed for each parameter to assess their relationship with the performance of the model. The number of LSTM Units, the quantity of neurons in the LSTM layer, was adapted ranging from 100 to 250. Diverse Batch Sizes (1, 2, 5, and 10) were applied in order to determine their effects on training stability and convergence. The model was trained for multiple Epochs of 200-500 each to ensure that the longer training duration does not impact the performance. A group of Optimizers including Adam, RMSprop, Nadam, and Adamax were implemented to enhance learning efficiency. Furthermore, Loss Functions such as MAE (Mean Absolute Error), MSE (Mean Squared Error), and Huber loss were adopted to decrease prediction errors. Besides this, the Feature Scaling techniques for example the StandardScaler and MinMaxScaler were used to normalize the data properly. The final part that presents the best parameters is saved for the results section of this document.

2.8 Proposed Model

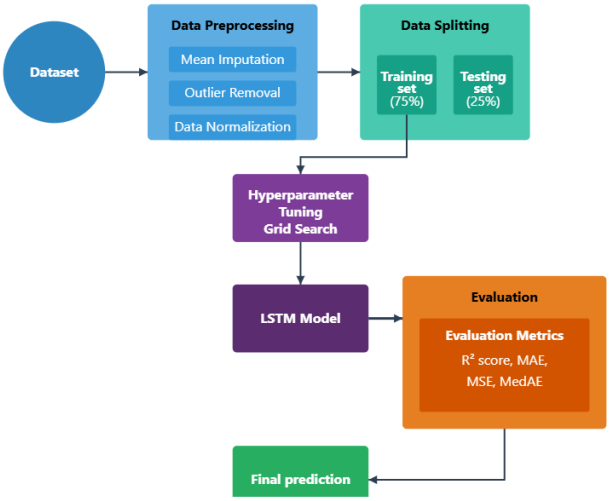

FIGURE 3. Flow chart of the proposed model

FIGURE 3 provides a diagrammatic representation of the proposed method. After loading the dataset, data preprocessing such as handling missing values through Mean Imputation, removing outliers through Interquartile Range and Data Normalization through MinMaxScaler, a feature scaling technique is performed to make the dataset better. The dataset was divided into training and testing set. 75% of data is taken for training and remaining 25% data for testing. Hyperparameter Tuning through Grid Search is performed on the pre-processed dataset to fine tune the LSTM model. The fine-tuned LSTM model with best parameter values is used to learn the training dataset and then predict the test dataset. Finally, the proposed model is evaluated using coefficient of determination (R2), Mean absolute

error (MAE), Mean squared error (MSE) and Median Absolute error (MedAE).

3. RESULTS AND DISCUSSION

3.1.Results

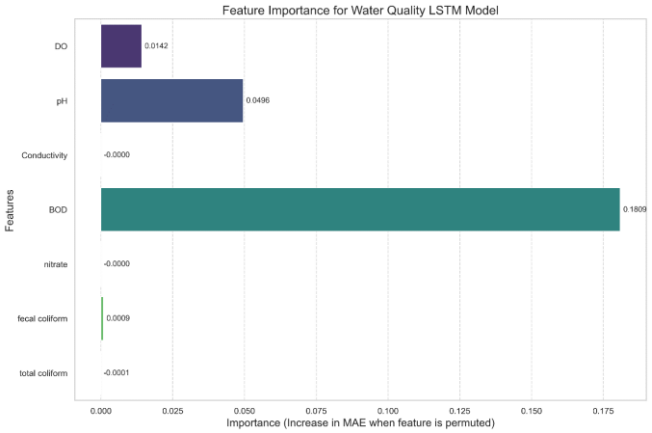

FIGURE 4. Feature Importance Analysis of Water Quality Prediction using LSTM.

As shown in **FIGURE 4**, during the Water Quality LSTM model test, Biochemical Oxygen Demand (BOD) emerged as the most significant feature scoring the highest among all of them, 0.1809.

Table 2. The settings of the best parameters using grid search algorithm

Parameters	Values Tried	Best Value
LSTM Units	100,150,200,225,250	200
Batch Size	1,2,5,10	1
Epochs	200,300,350,450,500	450
Optimizer	Adam, RMSprop,Nadam, Adamax	Nadam
Loss function	MAE, MSE, Huber	MAE
Feature Scaling	StandardScaler, MinMaxScaler	MinMaxScaler

To find the best parameters, this model used grid search approach. The details are shown in **Table 2**.

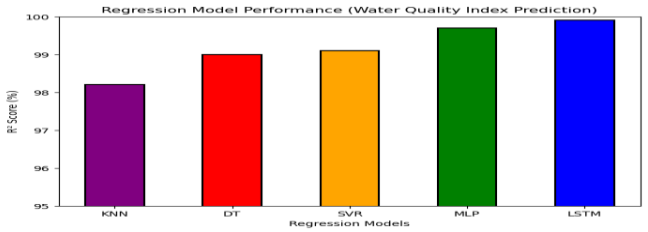

FIGURE 5. Comparison of R^2 score of LSTM and previous regression models

FIGURE 5 shows the predictive performance of different regression models that are used to predict the water quality index. It can be observed that the lstm has outperformed all the other models.

Table 3. Comparing LSTM with previous regression models

Models	MAE	MSE	MedAE	R^2Score
KNN regressor	0.009	0.002	0.005	98.25%
DT regressor	0.005	0.001	0.0013	99%
SVR	0.004	0.001	0.0012	99.1%
MLP regressor	0.003	2.8×10^{-5}	0.0009	99.8%
LSTM	0.0008	1.09×10^{-6}	0.0007	99.9%

Table 3 shows the results of the evaluation metrics of various models that were used to predict water quality index.The result shows us that the LSTM model outperforms every other model by achieving the least values for MAE (0.0008), MSE (1.09×10^{-6}), MedAE (0.0007) and highest value for R^2 score (99.9%) by proving its outstanding accuracy and precision.

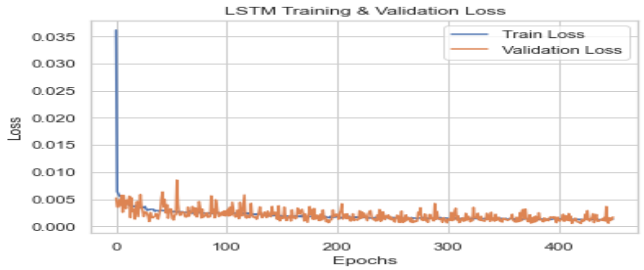

FIGURE 6.LSTM Training and Validation loss

In **FIGURE 6**, the beginning of the training loss depicts a sharp decline and it stabilizes at a lower value. The validation loss fluctuated and remained constantly low throughout.

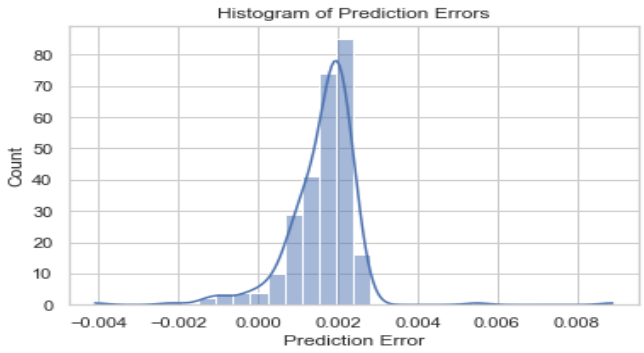

FIGURE 7. Prediction Error

FIGURE 7 shows the distribution of prediction errors. The errors are found to be distributed around zero.

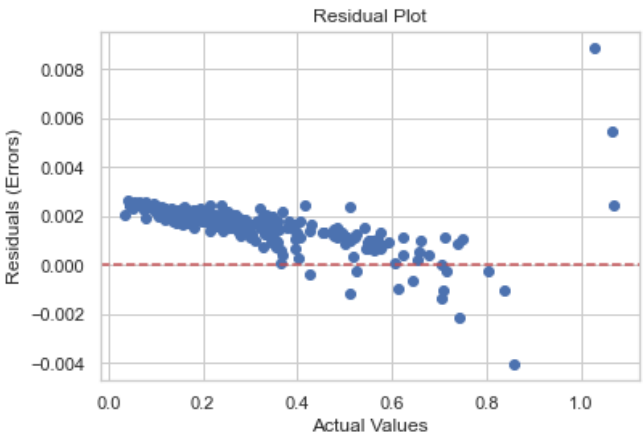

FIGURE 8. Residual plot

FIGURE 8 shows the difference between the predicted values and the actual values. It can be seen that most of that most of the residuals in this plot are close to zero and a few of them are way off.

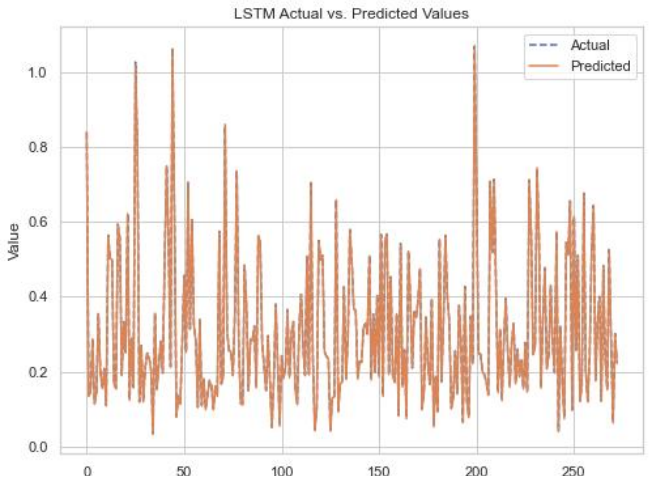

FIGURE 9.LSTM Model: Actual vs Predicted Values

In **FIGURE 9,** the similarity of the actual values and the predicted values can be seen through the graph.

3.2.Discussion

The proposed model's performance is compared to numerous existing models. To assess the effectiveness of the regression model, Mean Absolute error (MAE), Mean Squared error (MSE), coefficient of determination (R^2) and Median Absolute error (MedAE) were used.

Mean Absolute Error measures the average absolute difference between actual (y_i) and predicted ($\widehat{y_i}$).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

MAE of the model LSTM is:
0.000814303747241706

Mean squared error is used to measure the average squared difference between the predicted values and the actual values.

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

MSE of the model LSTM is:
1.094985494617778e-06

The coefficient of determination is a number between 0 and 1 that measures how well a statistical model predicts an outcome.

Table 4. Interpretation of Coefficient of Determination (R^2)

COEFFICIENT OF DETERMINATION(R^2)	INTERPRETATION
0	The model does not predict the outcome
Between 0 and 1	The model partially predicts the outcome.
1	The model perfectly predicts the outcome

R^2 Score of the model LSTM is:0.9999712426975482.

Median Absolute Error is the median difference between the observations (true values) and model output (predictions).

$$MedAE = median|y_{i.pred} - y_{i,true}|$$

MedAE of the model is:0.0007391101368321 44.

FIGURE 4 shows the feature important analysis. It suggests that the model is dependent on BOD when predicting water quality. pH (0.0496) and Dissolved Oxygen (DO) (0.0142) are contributory, but their contribution is relatively minor. Other features like fecal coliform (0.0009), conductivity, nitrate and total coliform (~0.0000 or negative values) are incorporated by the system but have inconsequential contribution to the prediction. This supports the assertion that BOD, pH, and DO - the three features underpinning the majority of BOD assessment organic pollution - have a low value toward improving model prediction. It is logical to enhance model performance by removing uninformative predictive features. So, the results suggest that the model accuracy could be improved by increasing focus on other important features while decreasing focus on less important features, which will likely enhance the efficiency of the model.

Table 2 details the tuning parameters investigated as well as the precise parameter values that resulted in the optimum performance based on the tuning.

Best Values for each values:

LSTM units - 200
Batch Size – 1
Epochs – 450
Optimizer – Nadam
Loss function – MAE
Feature Scaling - MinMaxScaler

These best parameters are very important in optimizing the model.

FIGURE 5presents the bar graph that compares the coefficient of determination of the LSTM model and the previous regression models. It demonstrated that deep learning models outperformed the classical ones. Hence, the LSTM model provided the utmost accuracy in the study on the Water Quality Index credentials, based on the R^2 score. Following closely is the MLP model with informative approximate values. Among the ordinary machine learning schemes, Support Vector Regression (SVR)had relatively sound outputs, scoring somewhat better than Decision Tree. In sharp contrast, K-nearest neighbors generated the least R^2scores for the WQI and failed grossly in providing the assumption conditions for generalization.

Table 3 compares the performance of LSTM with various other regression models such as K-nearest neighbors (KNN), Decision Tree (DT), Support Vector Regressor (SVR) and Multi-level Perceptron (MLP), using evaluation metrics such as Mean absolute error (MAE), Mean squared error (MSE), Median absolute error (MedAE) and coefficient of determination (R^2). It is evident that the LSTM model has least values for MAE, MSE and MedAE and highest value for R^2 score. It means that the LSTM model nearly explains all the variance of the target variable. Overall, the results indicate that the LSTM being most effective and reliable model for making predictions compared to traditional machine learning models making LSTM, the optimal model for this dataset.

As seen in **FIGURE 6**, the LSTM training and validation loss depicts performance with respect to the epochs. The y-axis displays the loss as ‘Mean Absolute Error’ and the x-axis displays the number of epochs. The beginning of the training loss depicts a sharp decline which means that the model learns quickly from the data. It also stabilizes at a lower value suggesting effective learning. The validation loss was low throughout and although it fluctuated at the start, it remained consistently low. This means that the model was able to generalize well without overfitting too much. The small gap between training and validation loss confirms that the model suffers little from high variance. Overall, the LSTM model is able to capture the water quality dataset while generalizing well which can be seen in the results of the graph.

FIGURE 7 shows the distribution of prediction errors reveals significant information concerning model accuracy depending on the frequency of error. The model shows good accuracy as the error is found to be distributed around zero, implying the deviations from the actual values are very small. The pattern is demonstrative of a zero-centered distribution, which stretches into reductions of greater magnitudes. The output corroborates the MAE results with the claim that the predictive model exhibits greater accuracy value estimation. The model's accuracy claim and his capture gap prove the model's reality limitation validity and reliability when looking at the small error dispersion.

FIGURE 8 depicts the difference between the predicted and actual values, enabling one to check how well the model is performing. Residuals should be randomly scattered around the red dashed line at zero, indicating that there is no visible pattern in the errors. Most of the residuals in this plot are close to zero, indicating that the model is making good predictions. But a few of them are way off, particularly at higher actual values, which could be where the model makes slightly incorrect predictions.

Overall, the plot shows that the model is well-fit with hardly any errors and no major bias in predictions.

FIGURE 9 is the plot of the actual values (dashed line) and predicted values (solid line) generated by the LSTM model against the values predicted by the model. The similarity of the actual and predicted values shows the ability of the model to predict by recognizing patterns in the data. The lack of divergence between the two shows very good predictive ability, and it is a sign that the LSTM model learns and generalizes well the pattern of the datasets.

CONCLUSION

The main aim of this study was to evaluate the performance of deep learning modelLSTM(Long Short Term Memory) in the prediction of surface water quality.LSTM was betterprediction of water quality index rather than other models like KNN,DT,SVR and MLP. The dataset used for the model constructions was “Indian water quality data”. The chemicalparameters were used for the proposed model as the input attribute. Furthermore, the modelexhibits less approximation of water quality index with less environmental or ecologicalvalues, Future studies aimed at the prediction of water index quality for Tamil Nadu surfacewater which are very much need in time and also to find the type of industries which affectsmore the water quality index.

REFERENCES

[1] Babu, C. N., & Reddy, B. E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data. Applied Soft Computing, 23, 27-38.

[2] Zhou, J., Wang, Y., Xiao, F., Wang, Y., & Sun, L. (2018). Water quality prediction method based on IGRA and LSTM. Water, 10(9), 1148.

[3] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210.

[4] Khan, M. A., Rahman, A., & Ali, S. (2021). IoT-based smart water quality monitoring: Recent developments and future directions. Sensors, 21(5), 1671.

[5] Ahmed, M., Mumtaz, R., & Hassan Zaidi, S. M. (2021). Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam, Pakistan. Water Supply, 21(6), 3225-3250.

[6] Rana, R., Kalia, A., Boora, A., Alfaisal, F. M., Alharbi, R. S., Berwal, P., ... & Qamar, O. (2023). Artificial intelligence for surface water quality evaluation, monitoring and assessment. Water, 15(22), 3919.

[7] Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 83(12), 35307-35334.

[8] Abbas, F., Cai, Z., Shoaib, M., Iqbal, J., Ismail, M., Alrefaei, A. F., & Albeshr, M. F. (2024). Machine learning models for water quality prediction: a comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan. Water, 16(7), 941.