Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

LSTM DRIVEN WATER QUALITY FORECASTING IN UPPER INDUS BASIN

Shafi Ullah[1], Afed Ullah Khan[*2], Saqib Mahmood[3], Fayaz Ahmad Khan[4]

[1,4] *National Institute of Urban Infrastructure Planning, University of Engineering and Technology, Peshawar 25000, Pakistan*
[2,3] *Department of Civil Engineering, University of Engineering and Technology Peshawar (Bannu Campus), 28100, Bannu, Pakistan.*

[*2] afedullah@uetpeshawar.edu.pk

DOI: https://doi.org/10.5281/zenodo.17340736

Keywords
Water Quality parameter, LSTM, Swat River, Forecasting, Climate Change

Article History
Received: 16 July 2025
Accepted: 26 September 2025
Published: 13 October 2025

Corresponding Author: *
Afed Ullah Khan

Abstract
This study evaluates the predictive capability of the Long Short-Term Memory (LSTM) model in forecasting water quality parameters of the Swat River at the Chakdara monitoring station using historical data from 1963 to 2020. The dataset was divided into training (70%) and testing (30%) subsets, and the model was developed and optimized using TensorFlow and Keras in Python. Feature engineering and hyperparameter tuning were applied to improve predictive performance for five key water quality indicators: Sodium (Na), Magnesium (Mg), Bicarbonate (HCO_3), Sulfate (SO_4), and Sodium Adsorption Ratio (SAR). The LSTM model achieved excellent results, with an R^2 of 0.95 (training) and 0.86 (testing), mean squared error (MSE) of 0.003 (training) and 0.008 (testing), Nash-Sutcliffe efficiency (NSE) of 0.93 (training) and 0.86 (testing), and root mean squared error (RMSE) values of 0.05 and 0.09, respectively. These results highlight the LSTM model's effectiveness in capturing long-term dependencies in hydrological time series, making it highly suitable for water quality prediction. The study underscores the potential of LSTM for integration into environmental monitoring systems to improve the reliability of water quality forecasting and support sustainable water resource management.

INTRODUCTION

Effective water quality monitoring is fundamental for sustainable water resource management, directly influencing ecosystem health, public well-being, and agricultural productivity. In developing countries like Pakistan, financial and technological limitations hinder proper monitoring and management of river water systems, especially in large water bodies such as the Indus River [1]. Conventional monitoring techniques, though reliable, depend on physical sampling and extensive laboratory analysis, making them costly, time-consuming, and labour-intensive [2]. Machine learning (ML) and deep learning models have emerged as powerful, cost-effective alternatives to traditional approaches, capable of handling complex, nonlinear, and temporal dependencies in hydrological data. Among these, Long Short-Term Memory Network (LSTM) is very popular for time series forecasting which eliminates the vanishing gradient problems and captures long-memory dependence in the sequential data strings [3]. This makes LSTM very well fit for water quality time series where long-term temporal patterns are all important for prediction success [4].

Although scientists in Upper Indus Basin have studied the various models like Artificial Neural Network (ANN), Gene Expression Programming

ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

(GEP), Multi-Expression Programming (MEP) etc. for the prediction of water quality parameters such as Total Dissolved Solids (TDS), Electrical Conductivity (EC) and Dissolved Oxygen (DO), application of LSTMs for water quality prediction in the Upper Indus Basin is not much explored [5]. This study addresses this gap in the literature by using LSTM networks to predict important water quality parameters viz. Sodium (Na), Magnesium (Mg), Bicarbonate (HCO3), Sulphate (SO4), and Sodium Adsorption Ratio (SAR) in the Upper Indus Basin [6]. Main aim is to evaluate the performance of LSTM in generating a precise forecasting of water quality parameters to minimize the use of the conventional methods, which are more labor-intensive, and also in aid of data-driven sustainable management of water resources.

Novelty of the research is its exclusive use of LSTM deep learning architecture to predict the river water quality in the Upper Indus Basin, Pakistan where such models have never been studied before. Unlike prior studies, which used conventional statistics or light learning algorithms, this study also applies long short-term memory (LSTM) which has a stronger ability to capture dependency structures in longitudinal water quality datasets, which have a time span of several decades. By considering a range of key parameters (Na, Mg, HCO3, SO4 and SAR) at once, the study has created a holistic predictive framework that, based on basic model input data, can make robust and high-quality forecasts with similar computational ease. This new technique is not only an innovation in methodological development of hydrological modelling, but also offers valuable lessons for sustainable water resource management in data-poor areas.

Study area Description

The present study deals with the Upper Indus Basin (UIB) of Pakistan of the Hindukush Himalaya region between latitudes 33°40′ to 37°12′ N and longitudes 70°30′ to 77°30′ E as shown in the Figure 1. The basin runs from the head of the upper Indus River beyond Tarbela Dam, it's a distance of some 1,150 km and the overall drainage blogging for nearly 165,400 km^2 [7]. The UIB has diverse topography with high notched plateaus, as well as some of the highest mountains of the world including peaks of the Karakoram over 7000m. The region has fairly low annual precipitation which ranges between 100 and 200 mm [8]. A unique Bureau feature is the existence of such a large area of glaciers (18,500 km2). These glaciers are an important source of the water, as their seasonal meltwater is a major source of run-off for the Indus River system.

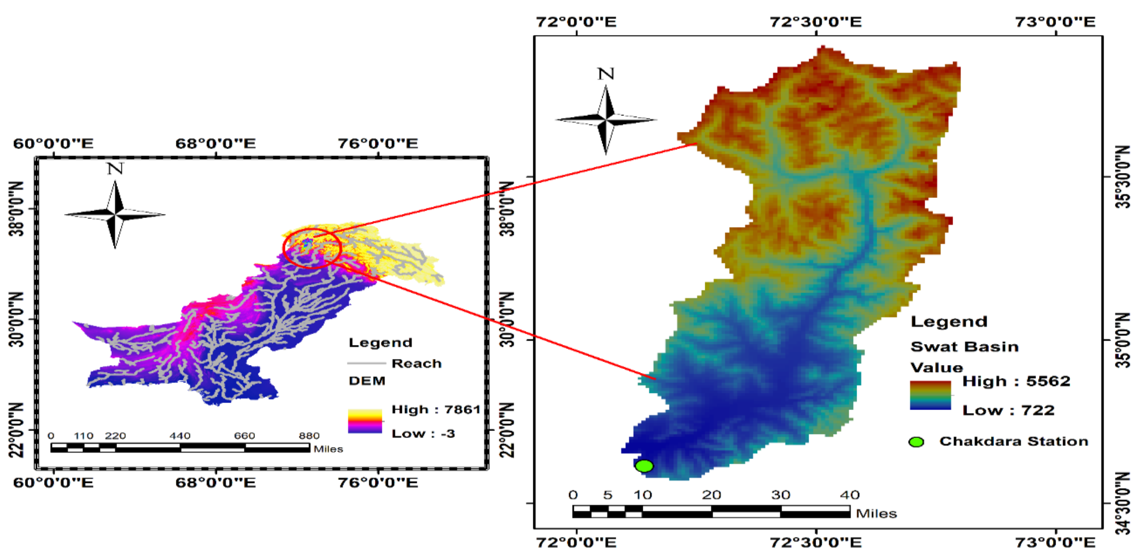

Figure 1: Study area map of the research

Data Collection

For the target study area, Digital Elevation Model (DEM) data was downloaded from the Earth Data portal of National Aeronautics and Space Administration (https://earthdata.nasa.gov/), and used for watershed delineation, with a spatial resolution of 30 m x 30 m, which was further utilized to represent the flood delineation analysis (Fig. 1). In addition, water quality data was retrieved from Water and Power Development Authority (WAPDA) monitoring station at Chakdara located in the Upper Indus River Basin. This dataset includes monthly chemical water quality and sediment monitoring results from 1963-2020.

Methodology

To evaluate the prediction performance of the Long Short-Term Memory (LSTM) model, the research workflow shown in Fig. 3 was followed, utilizing Python, Keras, and TensorFlow. Water quality data was obtained from the WAPDA monitoring station at Chakdara in the Upper Indus Basin. A data cleaning process was carried out to handle outliers and missing values. The raw dataset, which initially included 20 water quality parameters (e.g., Ca, Mg, Na, DO, pH, EC, carbonate, HCO_3, chlorine, sulphate, cations, anions, temperature, SAR, sand, silt, clay, Tmax, PCP, and PPM), was refined through feature selection. Using the Random Forest Algorithm, only parameters with significant correlations were selected to train the LSTM model.

The processed dataset was split into 75% for training and 25% for testing. Missing values in the dataset were replaced with the column-wise mean using:

$$x_{missing} = \frac{\sum_{i=1}^{n} x_i}{n}$$

To ensure uniformity across variables, the dataset was normalized between 0 and 1 using the min-max scaling technique. Feature engineering was then applied to enhance model performance by identifying the most significant predictors for the selected water quality parameters.

LSTM Model Development

The LSTM architecture was designed specifically to capture long-term temporal dependencies in the water quality time series data. The model consists of two stacked LSTM layers (128 and 64 units), dropout layers for regularization, and fully connected dense layers for output prediction (Fig. 2). The inputs to the model included key selected water quality parameters, while the outputs were the target variables Na, Mg, HCO_3, SO_4, and SAR.

The model was trained using the Adam optimizer and Mean Squared Error (MSE) loss function. Early stopping and model checkpoint call-backs were employed to prevent overfitting and improve generalization. Training was performed for up to 1500 epochs with a batch size of 16. The LSTM processed three inputs at each time step: the cell state, the hidden state from the previous step, and the current input vector. Through its internal gates (forget, input, and output), the LSTM updated the cell state, allowing it to preserve long-term dependencies and enhance predictive accuracy.

Hyperparameter Tuning

To optimize the LSTM model, hyperparameter tuning was performed by systematically adjusting:

- Learning rate
- Batch size
- Number of epochs
- Dropout rate
- Number of hidden layers and units
- Activation functions (ReLU, ELU, Sigmoid, Tanh)

Multiple training iterations were conducted until the model achieved maximum prediction accuracy with the best combination of hyperparameters.

Model Performance evaluation

The predictive performance of the LSTM model was evaluated using four widely adopted statistical metrics:

1. **Coefficient of Determination (R^2) :**

$$R^2 = 1 - \frac{\sum_{i-1}^{n} (O_i - P_i)^2}{\sum_{i-1}^{n} (O_i - \bar{O})^2}$$

2. **Nash-Sutcliffe Efficiency (NSE):**

$$NSE = 1 - \frac{\sum_{i-1}^{n} (O_i - P_i)^2}{\sum_{i-1}^{n} (O_i - \bar{O})^2}$$

3. **Mean Squared Error (MSE):**

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

4. **Root Mean Squared Error (RMSE):**

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

Where O_i is the observed value, P_i is the predicted value, $\bar{O}$ is the mean of observed values, and n is the number of data points.
Model performance was interpreted based on thresholds presented in Table 2, where values of $R^2 > 0.85$ and NSE ≥ 0.7 were considered very good, while low MSE and RMSE indicated minimal prediction error.
LSTM networks are a sort of RNN that is intended to address some of the shortcomings of regular RNNs, especially in dealing with long-term dependencies in sequential data [9]. Hochreiter and Schmidhuber [10] developed LSTMs in 1997, and they have since become one of the most popular and successful models for a range of sequence prediction problems. LSTMs make use of a complicated framework with gates to regulate information flow, overcoming the vanishing gradient problem and increasing network learning. The components of LSTM Model are the Forget Gate, which discards information, the Input Gate, which saves new information, the Cell State, which contains long-term information, and the Output Gate, which outputs information from the Cell State [11] as shown in Figure 2.

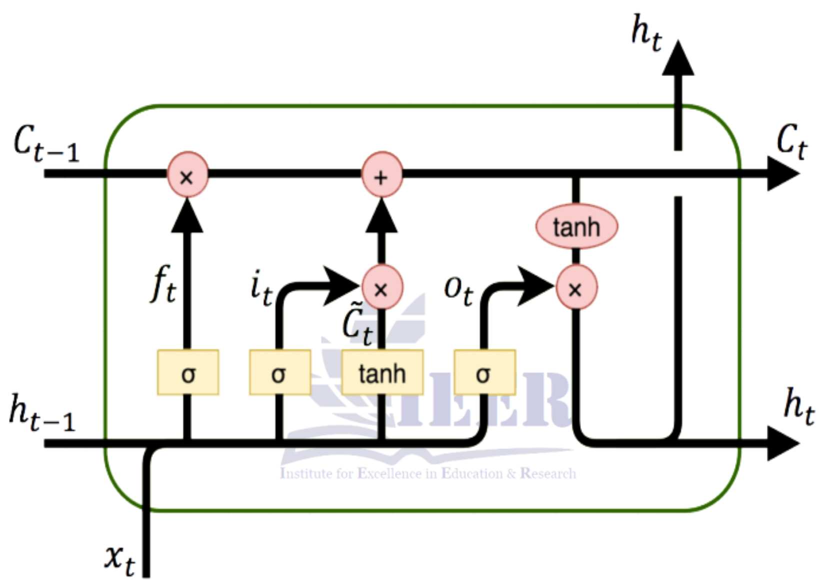

Figure 2: Typical LSTM architecture

LSTM model usually has three inputs, two input vectors come from the LSTM model itself and are generated by the LSTM model at the previous instant (t − 1). These inputs are shown as the cell state (C) and the hidden state (H) in figure 2. The third input vector "X" is provided from outside and is provided to the model at time instant "t". when the model is run by providing all three inputs the LSTM model processes the data through its internal gates and transforms the cell values of cell state and hidden state vectors [12]. The updated cell values will be the part of model inputs for next instant (t+1). Throughout this whole process of information flow control is done so that the cell acts as a long-term memory for next instant inputs.
By integrating forget, input, and output gates, LSTMs can update and maintain a cell state that spans long sequences and holds crucial information. Because of this, LSTMs are useful in many sequential data applications, such as time series prediction and language processing.
The flowchart of the methodology is shown in the Figure 3.

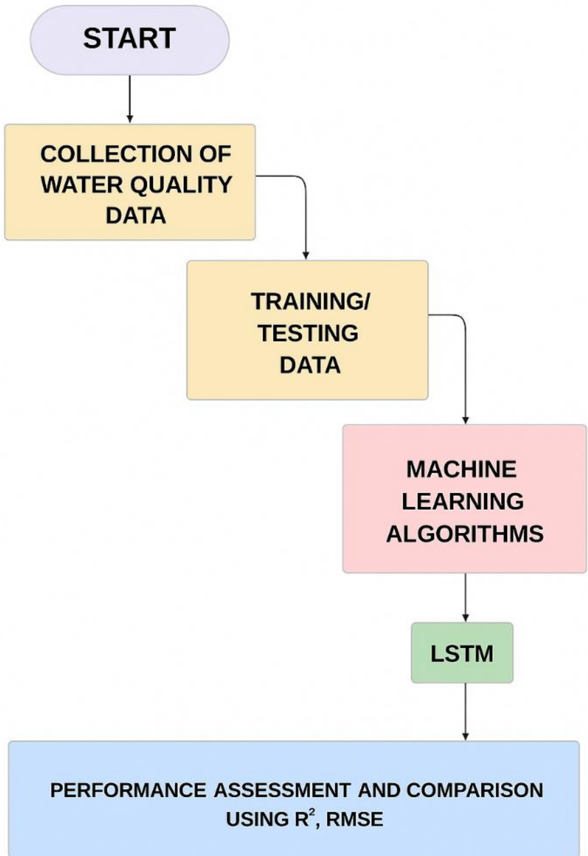

Figure 3: Methodology Flowchart

Results and Discussion

Long Short-Term Memory Networks (LSTM):

The LSTM model's performance, as shown in Table 1, demonstrates strong predictive capabilities for certain parameters, particularly Na and Mg. The model achieved an R^2 of 0.98 for Na during training and 0.87 during testing, indicating excellent predictive accuracy and generalization. Similarly, Mg exhibited a training R^2 of 0.91 and a test R^2 of 0.73, suggesting that while the model learns well, there is some drop in generalization. HCO_3 had a moderate training R^2 of 0.71 and a slightly lower test R^2 of 0.63, implying that the model struggles to fully generalize its learning. SO_4 had weaker test performance, with an R^2 of 0.40, indicating significant overfitting despite a decent training R^2 of 0.77. SAR achieved a fair level of training performance (R2 = 0.82), and training test performance was not reported and hence its generalization power is difficult to assess.

The reason might be the model performance in which one may see variations in performance depending on different factors like what type of optimizer, activation functions and hyperparameters were used. Adam was the most often used optimizer giving a reliable performance and rmsprop was used for SAR and maybe increasing the convergence of this parameter. The batch size and the number of epochs, also played an important part, such that the higher the epithet (1500), the better the training accuracy is going, in some points not really better on the test Notably, SO4 showed a lot of overfittings so it appears to be a good idea to use regularization techniques like dropout layers to achieve a better generalization result. Overall, while the LSTM model is good enough for capturing time-dependent patterns, there is still more to fine-tune in this model in order to make it that much better at generalization with respect to the model over all parameters.

ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

Table 1: Prediction Performance of LSTM Model

Parameter Name (meq/L)											
Target Variable	**Input Variables**	**Train / Test**	**R^2**	**MSE**	**NSE**	**RMSE**	**Optimizer**	**Activation**	**Batch Size**	**Epochs**	**Units**
Na	SAR, HCO3, Ca	Train	0.98	0.002	0.98	0.04	Adam	Relu	16	1500	128
		Test	0.87	0.007	0.87	0.08					
Mg	DS by Evap, Cl, EC	Train	0.91	0.012	0.91	0.11	Adam	Relu	32	1500	256
		Test	0.73	0.032	0.73	0.18					
HCO_3	DS by Evap, Ca, EC	Train	0.71	0.094	0.71	0.31	Adam	Relu	16	1500	128
		Test	0.63	0.049	0.63	0.22					
SO_4	DS by Evap, Na, SAR	Train	0.77	0.008	0.77	0.09	Adam	Relu	16	1000	128
		Test	0.40	0.019	0.40	0.14					
SAR	Na, Cations, Ca	Train	0.82	0.017	0.82	0.13	Rmsprop	Elu	16	100	128
		Test	0.82	0.008	0.82	0.09					

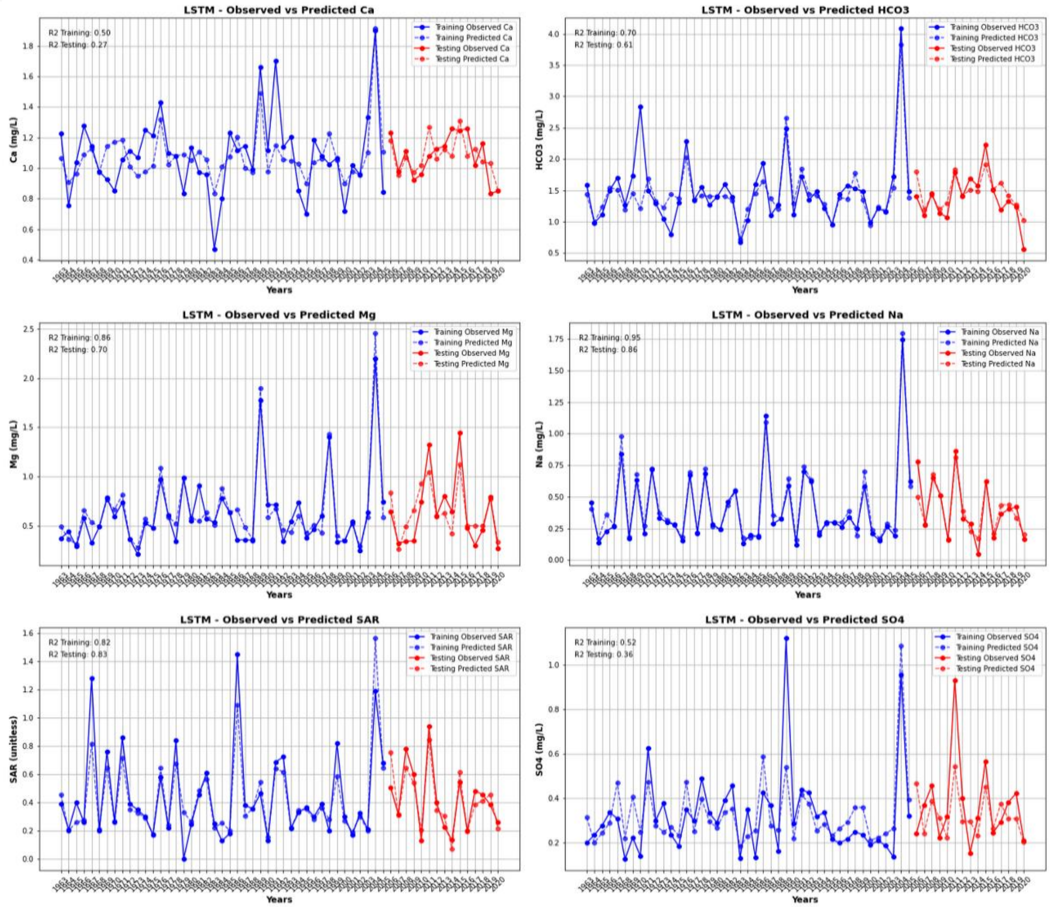

Figure 4: Model performance of the different parameter

Discussion

The findings from this study revealed that the LSTM model provided high accuracy of predictions of water quality parameters, by capturing well for relatively short time scale fluctuations as well as capturing long time-scale temporal dependencies. Better performance of LSTM is mainly due to its memory cell structure which allows the model to remember about relevant information over time, therefore LSTM can be considered as a more appropriate method for time-series forecasting problems, which is the water quality prediction problem in this case. The findings of the analysis confirmed the efficient use of LSTM by the incorporation of the nonlinear and dynamic features of water quality data by learning more complex temporality of water quality and preventing information loss of long-term context. This ability enables prediction of variations in key parameters with high confidence and high predictive skill of the model to offer useful information for environmental monitoring and management. Overall, the results illustrate the advantages of deep learning-based techniques (specifically LSTM networks) over statistical and shallow learning-based techniques in terms of forecast reliability. The success of LSTM employed in this study demonstrates the potential of this model as a tool for automating the water quality monitoring process, facilitating data driven water management and helping in sustainable environmental practices. LSTM is scalable and is a powerful tool for predictive environmental modeling with ability to learn from vast data-play and adapt to changing patterns of water quality.

Conclusion

This article shows the use of LSTM model is effective and capable of predicting the water quality parameters with high level of accuracy. The LSTM model was able to accurately represent long-term temporal correlations in the dataset and therefore predict the water quality primitive with better precision and reliability. The experiment shows that LSTM can be a very powerful model for environmental time-series prediction that offers better generalization ability and stability than the traditional models. The results demonstrate the expanding importance of deep learning for water resources management, especially for real-time monitoring and pollution hazard analysis. Additionally, LSTM-based forecasting can render more precise forecasts allowing for proactive choices-making, too early on detection of water quality decline, and for more sustainable water management strategies.

Future Work

Future studies may explore:

Hybrid architectures, such as CNN–LSTM or BiLSTM, to enhance feature extraction and temporal learning.

Incorporation of external variables (e.g., rainfall, temperature, land use, or human activity) to improve prediction performance.

Deployment of LSTM models in real-time monitoring systems integrated with IoT sensors for continuous water quality assessment.

REFERENCES

[1] S. Mahmood *et al.*, "Divergent path: isolating land use and climate change impact on river runoff," vol. 12, 2024.

[2] S. Khan *et al.*, "Intercomparison of SWAT and ANN techniques in simulating streamflows in the Astore Basin of the Upper Indus," vol. 88, no. 7, pp. 1847-1862, 2023.

[3] A. Ley, H. Bormann, and M. J. W. Casper, "Intercomparing LSTM and RNN to a conceptual hydrological model for a low-land river with a focus on the flow duration curve," vol. 15, no. 3, p. 505, 2023.

[4] F. Abbas *et al.*, "Machine learning models for water quality prediction: a comprehensive analysis and uncertainty assessment in Mirpurkhas, Sindh, Pakistan," vol. 16, no. 7, p. 941, 2024.

[5] S. A. Khaskheli, M. A. Rahu, S. Siraj, H. Jamshed, S. J. I. J. o. I. S. Iqbal, and C. Technologies, "Optimized water quality forecasting using machine learning," vol. 3, no. 2, pp. 46-60, 2024.

[6] M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. J. I. A. Anwar, "Toward design of internet of things and machine learning-enabled frameworks for analysis and prediction of water quality," vol. 11, pp. 101055-101086, 2023.

[7] M. Khan, A. U. Khan, S. Khan, F. A. J. W. S. Khan, and Technology, "Assessing the impacts of climate change on streamflow dynamics: A machine learning perspective," vol. 88, no. 9, pp. 2309-2331, 2023.

[8] S. Ali, D. Li, F. Congbin, and F. J. E. R. L. Khan, "Twenty first century climatic and hydrological changes over Upper Indus Basin of Himalayan region of Pakistan," vol. 10, no. 1, p. 014007, 2015.

[9] Y. Yu, X. Si, C. Hu, and J. J. N. c. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," vol. 31, no. 7, pp. 1235-1270, 2019.

[10] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory," vol. 9, no. 8, pp. 1735-1780, 1997.

[11] A. J. P. D. N. P. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," vol. 404, p. 132306, 2020.

[12] P. Fang, Y. Wang, Y. Zhao, and J. J. W. Kang, "Analysis of Prediction Confidence in Water Quality Forecasting Employing LSTM," vol. 17, no. 7, p. 1050, 2025.