Environmental Health Engineering and Management Journal. 2025;12:1472
http://ehemj.com
doi 10.34172/EHEM.1472
Environmental Health Engineering and Management Journal
Original Article
Open Access Publish Free

Predicting electrical conductivity with neural networks: A comparative study of ANN, RNN, CNN, and LSTM models

Muhammad Waqas[1*], Afed Ullah Khan[1], Jahanzeb Khan[2], Fayaz Ahmad Khan[3], Ateeq Ur Rauf[1]

[1]Department of Civil Engineering, University of Engineering and Technology Peshawar (Bannu Campus), 28100, Bannu, Pakistan
[2]Higher Education Department, Kohat 26000, Khyber Pakhtunkhwa (KP), Pakistan
[3]National Institute of Urban Infrastructure Planning, University of Engineering and Technology, Peshawar 25000, Pakistan

Abstract

Background: Electrical conductivity (EC) is an important indicator of surface water quality, primarily influenced by temperature, salinity, and human activities. The conventional EC experimental technique is resource-intensive and time-consuming. Recent advancements in machine learning provide an innovative technique for accurate EC prediction using historical time series data.

Methods: Surface water EC was assessed via four machine learning techniques, namely Artificial Neural Network (ANN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). The predictive capability of the aforementioned models was assessed via six statistical performance indicators, namely Coefficient of determination (R^2), Percent Bia (PBias), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Root Mean Square Error (RRMSE), and Nash-Sutcliffe Efficiency (NSE).

Results: The findings of the present research work show that the LSTM model outperforms in predicting EC. The LSTM model's efficacy was demonstrated by its outstanding R^2 values of 0.99 and 0.94 during training and testing, respectively. Notably, RNN, ANN, and CNN ranked second, third, and fourth, respectively, based on statistical performance indicators.

Conclusion: The results show that LSTM outperforms the remaining models in predicting EC. The findings of this study can assist water quality managers in finding the optimum machine learning model for modeling EC in the understudied area. Overall, this work advances our understanding of EC prediction using machine learning techniques.

Keywords: Water quality, Rivers, Prediction, Neural networks, Machine learning

Citation: Waqas M, Khan AU, Khan J, Khan FA, Rauf AU. Predicting electrical conductivity with neural networks: A comparative study of ANN, RNN, CNN, and LSTM models. Environmental Health Engineering and Management Journal. 2025;12:1472. doi: 10.34172/EHEM.1472.

Article History:
Received: 5 November 2024
Revised: 8 March 2025
Accepted: 10 April 2025
ePublished: 14 September 2025

***Correspondence to:**
Muhammad Waqas,
Email: engr.muhammadwaqas88774@gmail.com

Introduction

Water, as an essential resource, plays a significant role in sustainable life on the Earth's surface. It plays a significant role in agriculture, industries, sanitation, public health, etc. Access to safe and clean water is essential for human survival. However, nowadays, due to growing populations, rapid urbanization, increased industrial and agricultural activities, and untreated effluent discharges, surface water quality is degrading day by day. According to a survey conducted by the UNICEF and World Health Organization (WHO), approximately 2 billion people live in high water-stress countries worldwide, and around 2.2 billion people have no access to clean water. They are using unsafe drinking water, which is responsible for 485,000 diarrheal deaths annually, especially in developing countries (1). Pakistan, as a developing Asian country, is also facing challenges related to water quality pollution. Approximately 60% of Pakistan's rivers are heavily polluted, with harmful levels of chemical pollutants, including pesticides, heavy metals, and untreated sewage (2). This has far-reaching consequences for both ecosystems and public health. These alarming situations underscore the urgent need for water quality management and prediction. Water quality parameter testing is a time-consuming task that necessitates the use of specialized expertise and equipment (3). Conversely, machine learning-based water quality simulations provide a solid alternative to traditional laboratory testing. Various models, including deterministic, statistical, stochastic, and numerical models, are extensively used for simulating water quality. The aforementioned models, however, have several shortcomings, including a complex structure and a reliance on large amounts of data for model creation (4–7). Moreover, the aforementioned models had low forecast precision in predicting water quality. Statistical techniques for water quality modeling typically assume a linear relationship between dependent and independent variables.

Literature presents that the use of traditional approaches (manual sampling and laboratory tests) for water quality monitoring was widespread earlier (8,9). Although these methods were efficient in localized water quality monitoring, they were invaluable, yet time-consuming, and limited in their applicability to large and dynamic datasets (10). Furthermore, these methods were also unable to predict the real-time water quality parameters. Therefore, these were shifted to advanced models such as statistical and machine learning models.

Several studies in Pakistan have utilized statistical models (e.g., PCA, regression analysis, correlation analysis, Cluster analysis, and Water Quality Index) to assess the water quality of various rivers, including the River Swat, River Hunza, Indus River, and River Ravi, etc. (11–15). Despite their extensive use, they have some drawbacks, as they can only handle stationary and linear datasets (16). Conversely, AI models surpass the constraints of conventional models by being able to analyze non-linear data and complex hydrological and environmental processes. This motivated the authors to employ advanced AI techniques, as these models have the capability of predicting and modeling the internal link between water quality parameters and modeling their time series. Machine learning models, including gene expression programming (GEP), artificial neural networks (ANN), Hybrid RT-Artificial Neural Network (RT-ANN), Random Forest (RF), gradient boosting, polynomial regression, etc., have enhanced the accuracy in recent works (17–22). These models were able to capture non-linear data and deal with large datasets. Despite their superior performance in handling complex data as compared to the traditional and statistical models, these models still face challenges in dealing with time dependencies and time-series data. This led to the incorporation of advanced AI models, such as LSTM, ANN, RNN, and CNN, to improve accuracy and overcome the limitations of previous models. These models are efficient in handling complex time-series data and analyzing high-dimensional interactions between water quality parameters. The LSTM model is effective in time-series data and long-term dependencies, while ANN and RNN can manage nonlinear and dynamic datasets. Moreover, CNN can capture spatial correlation. The use of these advanced models enables accurate prediction, marking a significant advancement in water quality prediction as compared to other approaches. These properties of deep learning models inspired the authors to carry out the present research work.

This study aimed to compare the predictive capability of deep learning techniques in simulating electrical conductivity at the Bara River Basin. Initially, the model's predictive capability was assessed using statistical performance indicators. Later, the best-performing model was selected via compromise programming. The superior model will be used for managing the water quality of the study area.

The novelty of the present study lies in its comprehensive approach to predicting the electrical conductivity (EC) of river water by leveraging multiple neural networks, including ANN, RNN, CNN, and LSTM. Unlike previous studies that typically relied on a single model, this research integrates and compares multiple deep learning techniques to enhance prediction accuracy and reliability. Furthermore, the study presents the first application of neural network-based EC prediction for the river Bara, providing a pioneering assessment of advanced machine learning models. The research identifies the most effective approach, thereby contributing valuable insights for improved water quality management and environmental sustainability.

Study area

The Bara River, a prominent geographical feature in Pakistan's Khyber Pakhtunkhwa region, originates in the Tirah Valley in Bara Tehsil, Khyber Agency. The river's journey begins in the natural splendor of the Tirah Valley, a place distinguished by its distinctive terrain and cultural diversity. The Bara River flows from its source across a varied topography before joining the Kabul River Canal, whose waters originate from the Warsak Dam. The Bara River then continues north-eastward, flowing through the scenic Nowshera District. Its journey comes to an end when it enters the Kabul River at Camp Koruna in Akbarpura. The river has significant variations in discharge, with annual runoff ranging between 0.02 and 0.23 million acre-feet (MAF) (23). The total catchment area of the river is 1793 km^2 (24). The water quality monitoring station is located at Jhansi post, with a latitude of 33.8707 and a longitude of 71.4092. This area was chosen because it represents the traits and dynamics of the Bara River in Khyber Pakhtunkhwa, Pakistan. Figure 1 depicts the study area.

Materials and Methods

Data Collection and Description

The water quality data for the Bara River at Jhansi post station, spanning from 1963 to 2007, were collected from the Surface Water Hydrology Department and the Water & Power Development Authority (WAPDA). The dataset comprised Ca, Mg, Cl, HCO_3, Na, SO_4, Sodium Adsorption Ratio (SAR), Suspended solids, electrical conductivity, water temperature, dissolved oxygen, turbidity, and pH indicators. These aforementioned parameters play a vital role in water quality assessment.

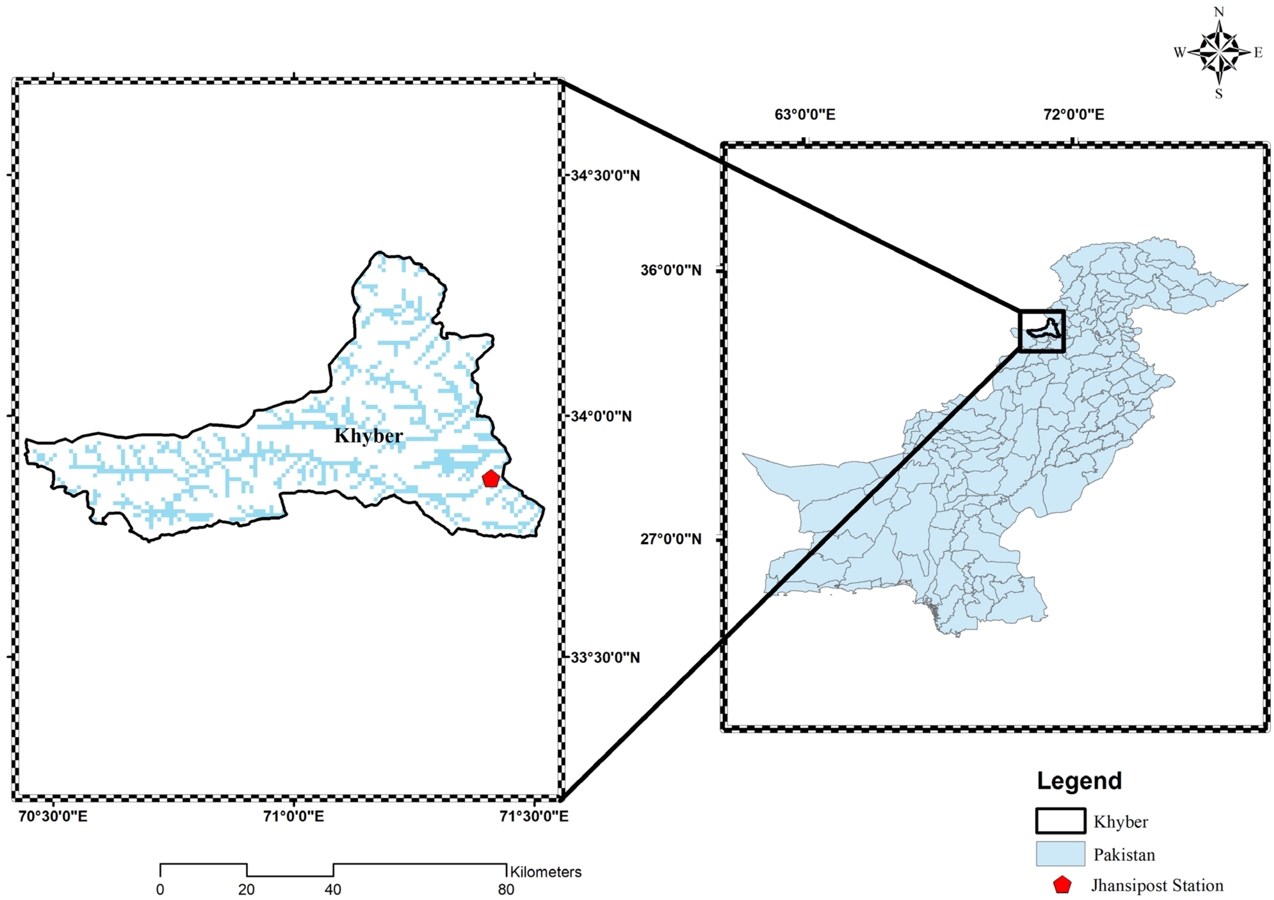

Figure 1. Description of the study area

Techniques Used

This study aimed to compare the predictive capability of deep learning techniques in simulating electrical conductivity at the Bara River Basin. Initially, the model's predictive capability was assessed using statistical performance indicators. Later, the best-performing model was selected via compromise programming. The step-by-step procedure used in this study is demonstrated Figure 2.

Data Preprocessing

Data preprocessing is the process of preparing raw data for analysis using various operations and transformations, including handling missing values and outliers, as the collected data included minor missing values, which were addressed by applying the Mean Imputation approach. Outliers are identified using the Interquartile Range (IQR) method. The values beyond 1.5 times the IQR 1st and 3rd quartiles were capped to the nearest valid range. Moreover, Min-max scaling is applied to the dataset before model training to standardize the water quality indicators. This stage ensured that all water quality indicators fell within a specific range, which improved the model's ability to mitigate the impact of dominant water quality indicators on model training. Min-max scaling is commonly used in neural networks and is effective in handling datasets without outliers (25). This approach has better performance in Artificial Neural Networks and also facilitates faster convergence.

Train-Test Split

The test-train split is the process of evaluating a model's performance, specifically how it performs with new data. In this research work, the observed water quality data were divided into two parts: training (80%) and testing (20%). The training dataset was used for model development, while the testing dataset was used for assessing generalization performance.

Deep learning techniques

Artificial Neural Network (ANN)

ANN is a machine learning model that works like a biological neuron network. It consists of interconnected units that work together to solve a complex problem. It is used for both type analysis, including classification and regression. In the present study, the ANN model consists of three layers: an input layer, a hidden layer, and an output layer, as shown in Figure 3.

These layers form a network of interconnected nodes, which serve as the foundation of this model. During the training process, the networks adjust the assigned

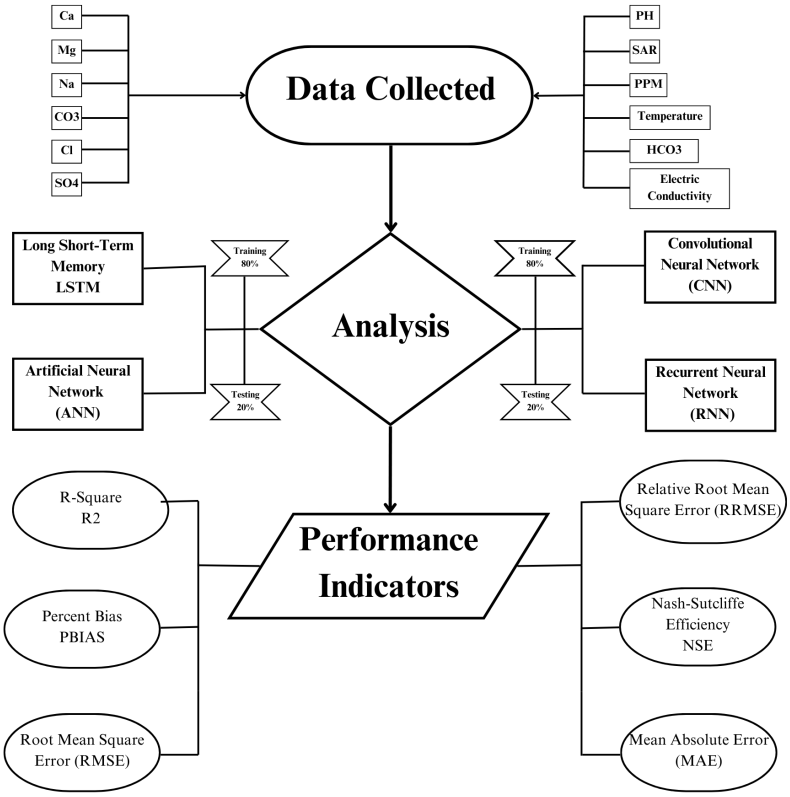

Figure 2. The methodology of the study

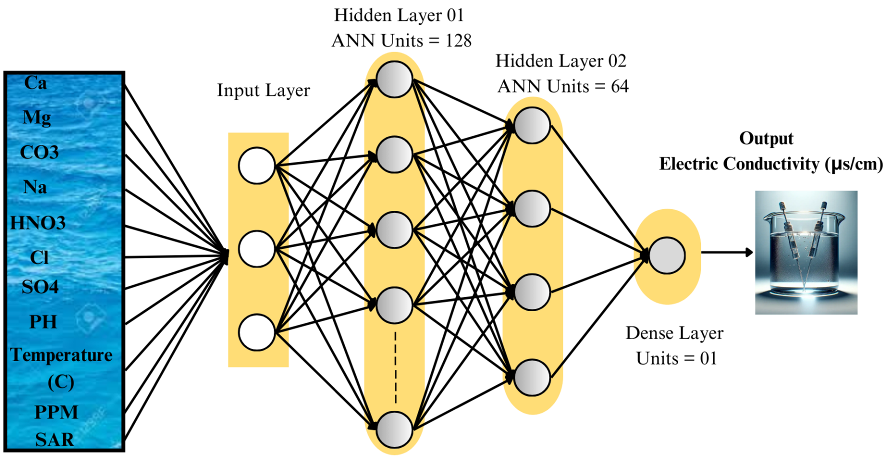

Figure 3. Architecture of the ANN Model

weightage to learn the inherent pattern in water quality data. The activation function of all nodes uses the inputs as a weighted sum to generate the output. The ANN model can be applied to various purposes, including time series prediction. When dealing with inherent temporal dependencies and long-term memory, more sophisticated methods like RNN and LSTM are required. Mathematically, the model is divided into two parts, including the input to hidden and the hidden to output.

Input to hidden state:

$$Output\ from\ hidden\ layyer = H = f\left(W^{1}.X + b^{(1)}\right) \tag{1 a}$$

Where f is the activation function for the hidden layer, W^{1} is a weighted matrix for input to the hidden layer, $b^{(1)}$ is the bias vector for the hidden layer, and X is the input vector.

Hidden to output state:

$$Output\ vector = f\left(W^{2}.H + b^{(2)}\right) \tag{1 b}$$

Where f is the activation function for the output layer, W^{1} is a weighted matrix for the hidden to the output layer, $b^{(1)}$ is the bias vector for the output layer, and H is the output from the hidden layer.

Recurrent Neural Network (RNN)

The RNN model has loop-type connections. It allows the layers to remain hidden, which acts as a memory for earlier input data. This earlier input data enables the model to extract information from previous points in the sequence. Due to its memory-preserving ability, the RNN model performs well in applications such as time series predictions. The RNN model was introduced to address the drawbacks of traditional feedforward neural networks while dealing with sequence data. The mathematical equation of the RNN model is given by.

Hidden state:

$$ht = \sigma\left(W_{hh}h_{t-1} + W_{xh}x_t + b_h\right) \quad (2\text{ a})$$

where ht is the result of the hidden state, W_{hh} is a weighted matrix from the previous time step, h_{t-1} is hidden state from the previous step, W_{xh} is a weighted matrix for the input at the current time step, x_t is input for the current time step, b_h is a bias vector for the hidden state, and σ is an activation function for the hidden state.

Output calculation:

$$y_t = \Phi\left(W_{hy}h_t + b_h\right) \quad (2\text{ b})$$

Where y_t is output at the time step t, Φ is the activation function for the output, W_{hy} is a weighted matrix from the hidden to the output state, and b_h is a bias vector for the output state.

Long Short-Term Memory (LSTM)

LSTM is an advanced form of RNN model, introduced to solve the long-term dependencies problem. Due to self-connected recurrent pathways, LSTM can store and capture information from long sequences. It performs well in predicting time series data due to its robust architecture, which captures complex temporal relationships. Accurate water quality prediction mainly depends on observed data patterns. LSTM performs well where past observations and context have a significant influence on future predictions. Mathematically, it can be expressed as follows.

Forget State: $f_n = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right)$ (3 a)

Input Gate: $i_t = \sigma_g\left(W_i x_t + U_i h_{t-1} + b_i\right)$ (3 b)

Output Gate: $O_t = \sigma_g\left(W_o x_t + U_o h_{t-1} + b_o\right)$ (3 c)

Output Gate: $C'_t = \sigma_g\left(W_c x_t + U_c h_{t-1} + b_c\right)$ (3 d)

Cell State: $C_t = f_t c_{t-1} + i_t C'_t + b_c)$ (3 e)

Cell State: $h_t = O_t \sigma_{t-1}\left(C_t\right)$ (3 f)

Here W_f, W_i, W_o, W_c, U_f, U_i, U_o, U_c are weighted matrices and b_f, b_i, b_o, b_c are biases.

Convolutional Neural Network (CNN)

CNN is a deep learning algorithm designed for the automatic extraction of spatial features from structured data. It consists of multiple layers, including input, hidden (convolutional and pooling), and output, which detect the complex nature of data, learn features, and make predictions based on these features. It can be applied in various fields, including time series analysis, although image analysis is still its most common application. Time series forecasting tasks in water quality research can benefit from CNNs because they provide an appropriate balance between interpretability and model complexity. The CNN model can be mathematically shown as follows.

Convolutional Operation:

$$z_{i,j,k} = \sum_{m=0}^{M-1} . \sum_{n=1}^{N-1} . \sum_{c=0}^{C-1} x_{i+mj+n,c} \; . w_{m,n,c,k} + b_k \quad (4\text{ a})$$

Where $Z_{i,j,k}$ is the output feature map at position (I, j) for the k-th filter, $x_{i+mj+n,c}$ is the input value at position (i, mj + n) in the c-th channel, $w_{m,n,c,k}$ is the weight of the filter at the position (m, n) in the c-th channel for the k-th filter, b_k is a bias term for the k-th filter, M x N is the size of the filter, and C is the number of input channels.

Activation function:

$$a_{i,j,k} = ReLU\left(z_{i,j,k}\right) = \max\left(0, z_{i,j,k}\right) \quad (4\text{ b})$$

where $a_{i,j,k}$ is activation output at position (I, j) in the k-th feature map and $z_{i,j,k}$ is the convolution output at position (I, j) in the k-th feature map.

Pooling Operation: $p_{i,j,k} = max_{(m,n)\in PxQ}\left(a_{i,mj+n,k}\right)$ (4 c)

where $p_{i,j,k}$ is the pooled output at position (I, j) in the k-th feature term.

Fully connected layers: $y_I = \Phi(\sum_k p_k . \; w_{I,k} + b_I)$ (4 d)

Where y_I is the output of the I-th neuron in the fully connected layer, $w_{I,k}$ is a weighted value for connecting k-th input to I neurons, b_I is bias term for I-th term, and p_k is the flattened pool value.

Output layer: $\acute{y} = \frac{exp(yi)}{\sum_{j=1}^{K} exp(yi)}$ (4 e)

where ý is the predicted probability for i-the class, yi is input to the softmax function for the i-th class, and K is the number of classes.

Neural Network Architecture and Optimization

In this study, ANN, RNN, LSTM networks, and CNN were implemented using Python, utilizing the Keras and TensorFlow libraries. The architectures of these networks were carefully designed and optimized through a trial-and-error approach to determine the best-performing model for water quality prediction.

ANN

The ANN model consisted of an input layer, two hidden layers with 128 and 64 units, and an output layer employing the ReLU activation function Figure 3. The Adam optimizer was used to optimize the network, while the MSE loss function was applied to evaluate model performance.

RNN

The RNN model followed a similar structure, comprising an input layer, two recurrent hidden layers with 128 and 64 units, and an output layer Figure 4. The Adam optimizer and MSE loss function were employed to optimize and assess the model's performance.

LSTM Network

The LSTM model was developed for regression tasks and consisted of an input layer, two LSTM layers, and an output dense layer Figure 5. The first LSTM layer had 128 units, followed by a 64-unit second LSTM layer with ReLU activation. To prevent overfitting, dropout regularization (0.2) was applied between the LSTM layers. The model was optimized using the Adam optimizer and evaluated with the MSE loss function.

CNN

The CNN model was tailored to capture spatial-temporal patterns in water quality data using Python Figure 6. It comprised three main layers: input, hidden, and output layers. The hidden layers included a convolutional layer with 256 units and a pooling layer with 128 units, designed to recognize high-order dependencies among water quality variables. By leveraging hierarchical patterns, the CNN effectively captured relationships between water quality parameters and improved prediction accuracy.

Optimization Strategy and Model Performance

The optimal architectures of the networks were achieved through a trial-and-error approach, as a method commonly practiced in deep learning studies. Various hyperparameters, including the number of layers, number of neurons per layer, activation functions, and dropout rates, were adjusted iteratively to enhance model accuracy.

For all models, the Adam optimizer was employed with a learning rate of 0.001, and Min-Max scaling was used for data normalization. The dataset was split into 80% training and 20% testing subsets. The models' predictive capabilities were evaluated using statistical performance indicators, which demonstrated their effectiveness in predicting observed water quality data.

Evaluation Metrics

The statistical performance metrics, namely the coefficient of determination (R^2), Root Mean Square Error (RMSE), Relative Root Mean Square Error (RRMSE), Mean Absolute Error (MAE), Percent Bias (PBIAS), and Nash-Sutcliffe Efficiency (NSE), were used for evaluation of the models (26).

R^2 is a statistical indicator used in regression and other statistical models for assessing the variance in observed and predicted data. It indicates how well a model captures and predicts the observed data. Its value ranges between 0 and 1, where a higher value shows the best fit. RMSE is used to measure the error between the observed and modeled values. Its value varies from 0 to positive infinity. A low RMSE indicates that the model fits well and predicts the observed data more accurately. RRMSE assesses the model's predictive capability based on relative errors. MAE assesses model predictive capability based on the absolute error between observed and model data. PBIAS is a statistical indicator used for assessing overestimation and underestimation. Its value ranges between negative infinity and positive infinity. The value closer to indicates the best fit. NSE is widely used to assess model

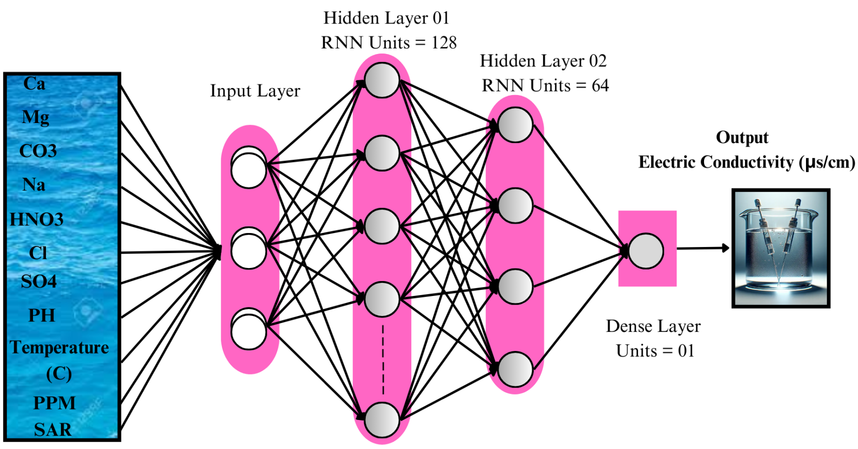

Figure 4. Architecture of the RNN Model

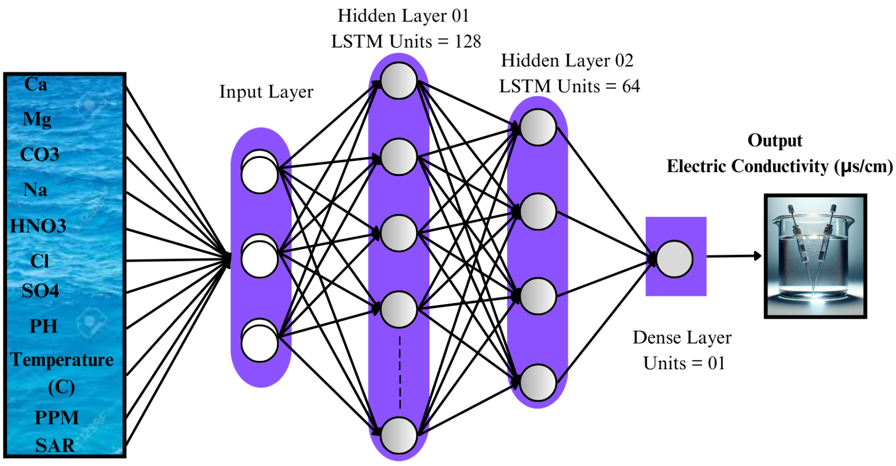

Figure 5. Architecture of the LSTM Model

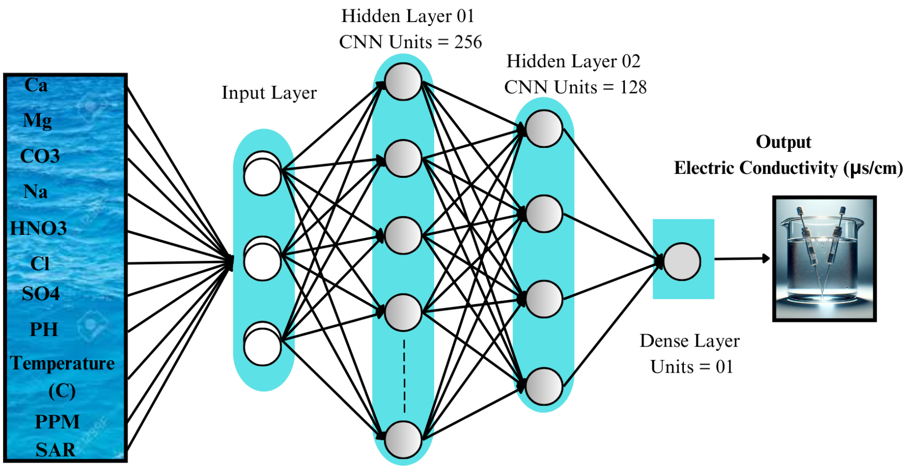

Figure 6. Architecture of the CNN Model

performance in hydrological studies. NSE value ranges between $-\infty$ and 1, where a value of 1 indicates the best fit.

Compromise Programming (CP)

CP is a well-established technique, which is based on statistical performance metrics (27). It ranks statistical models based on their overall performance, rather than considering a single statistical indicator. In this research, CP was used to rank deep learning models. Statistical performance indicators, namely R^2, RMSE, RRMSE, PBIAS, MAE, and NSE, combinedly measure the effectiveness of deep learning models in predicting the observed electrical conductivity. Using the aforementioned statistical performance indicators, a specific distance metric termed as (LP) was computed in the context of CP (28). LP quantifies the distance between the actual model performance values and the ideal values (where the model perfectly predicts a target). The model having the lowest LP value is considered the best and optimal choice. The LP equation is as follows.

$$\mathrm{Lp} = [\sum_{n=1}^{n} |Wn* - Wn|^{m}]^{\wedge}1/m$$

Where Wn denotes the observed value of a statistical performance measure, and Wn* denotes the ideal value of the performance measure, obtained when the modeled output resembles the observed data. m is a parameter (typically $m=1m=1$ or $m=2m=2$) that controls the sensitivity of the metric to deviations from the ideal value.

Results

In this study, the predictive capabilities of neural networks were compared based on their ability to predict electric conductivity. Figure 7 depicts the density and the summary statistics of each water quality parameter. The collected data, shown in Figure 7, are divided into two parts: training (80%) and validation (20%), and were analyzed using deep learning models. All the models performed best in the training phase, as shown in Table 1, but were unable to maintain the same performance in the testing phase.

The ANN exhibited good predictive performance across various configurations. The ANN model achieved R^2 values of 0.99 in training and 0.61 in testing, using the Adam optimizer with a learning rate of 0.001 and the

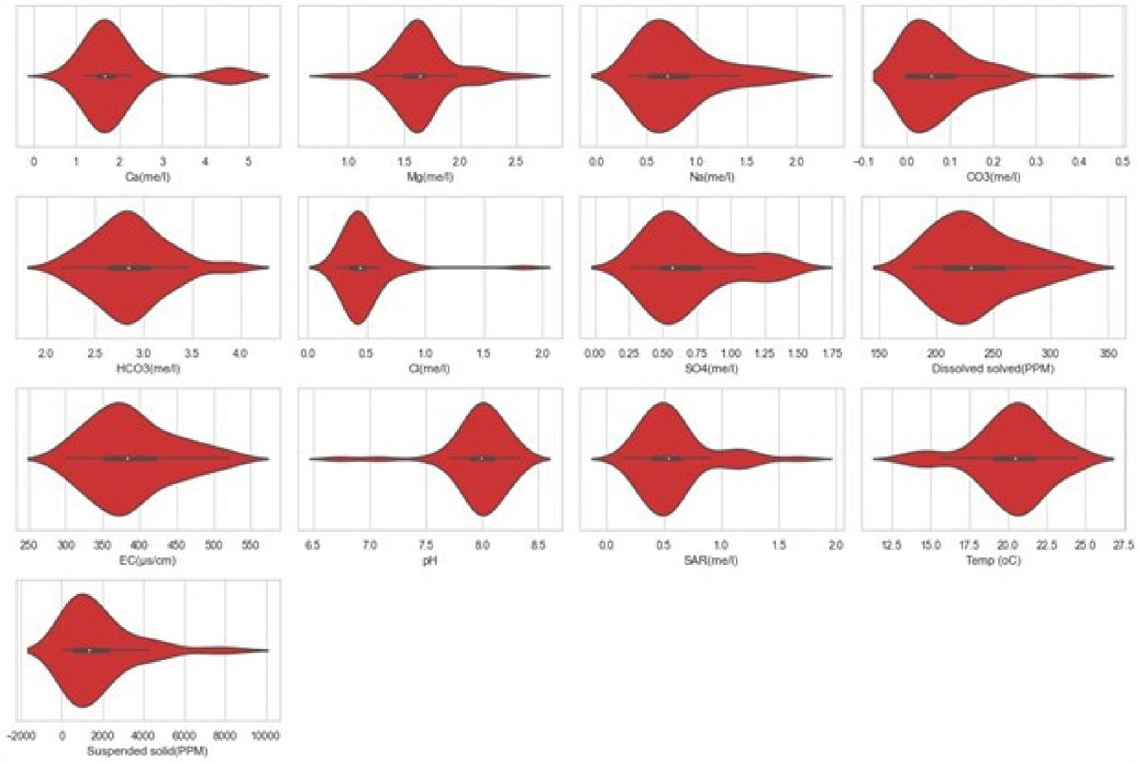

Figure 7. Violin plots of water quality parameters(29)

Table 1. Comparison of Neural Networks

Model	Training						Testing					
	R^2	RMSE	RRMSE	PBias	MAE	NSE	R^2	RMSE	RRMSE	PBias	MAE	NSE
ANN	0.9999	0.00000004	0.00000016	-0.00000045	0.00000003	0.999999	0.614569	0.2116272	0.620830	8.84698	0.155782	0.614569
RNN	0.9998	0.002873	0.01308866	0.093960	0.001964	0.999828	0.856386	0.1291801	0.378963	-10.16024	0.104640	0.856386
LSTM	0.9974	0.011122	0.05065811	-0.025626	0.006209	0.997433	0.944757	0.0801188	0.235036	-7.562904	0.066148	0.944757
CNN	0.9921	4.284229	0.01110000	-0.000600	3.000039	0.992126	0.938567	16.708400	0.041400	-0.004200	12.38479	0.950319

Min-Max scaling technique (30,31). The higher R^2 in the training phase reveals that the model closely fits the data pattern and precisely predicts the data in the training phase. However, the notable drop in R^2 during the testing phase indicates that, due to the complex nature of the data, the model was unable to capture the data pattern and failed to maintain its superior performance during the testing phase. It was also noted that model accuracy increases with the increasing number of epochs. Since the ANN model is sensitive to the number of epochs, hyperparameter adjustment is crucial for achieving good results. When it came to forecasting electrical conductivity in surface water, the ANN proved to be a good model, laying the groundwork for future research. The observed and predicted data are visualized using a line plot, which separates the training and testing data with a boundary line, as shown in Figure 8 (a).

Figure 8 (b) shows the results of the RNN model. The predicted values line is very close to the observed line in both training and testing phases, showing the good performance of RNN in predicting electrical conductivity. The RNN model achieved an R^2 value of 0.99 during training and 0.85 during testing, utilizing the Adam optimizer with a learning rate of 0.001 and the Min-Max scaling approach (32). The higher R^2 indicates that the model learned has the data pattern and can precisely predict the data in the training phase. However, the significant drop in R^2 during the testing phase indicates the model's underperformance. The underperformance of the ANN model in the testing phase is due to its inability to capture the patterns in time series data. Moreover, due to the vanishing gradient problem, the RNN model forgets information more quickly than the LSTM, which makes it unsuitable for handling long-term data. However, it is still evident that RNN is good at capturing the temporal dependencies of data. Particularly, the RNN model overweighs the ANN technique in terms of R^2, especially in the testing phase, indicating its capability in water quality time series prediction. Interestingly, the model's accuracy was not significantly improved by adding more epochs. The findings of this study reveal that the RNN model is adept at analyzing sequential data, which may

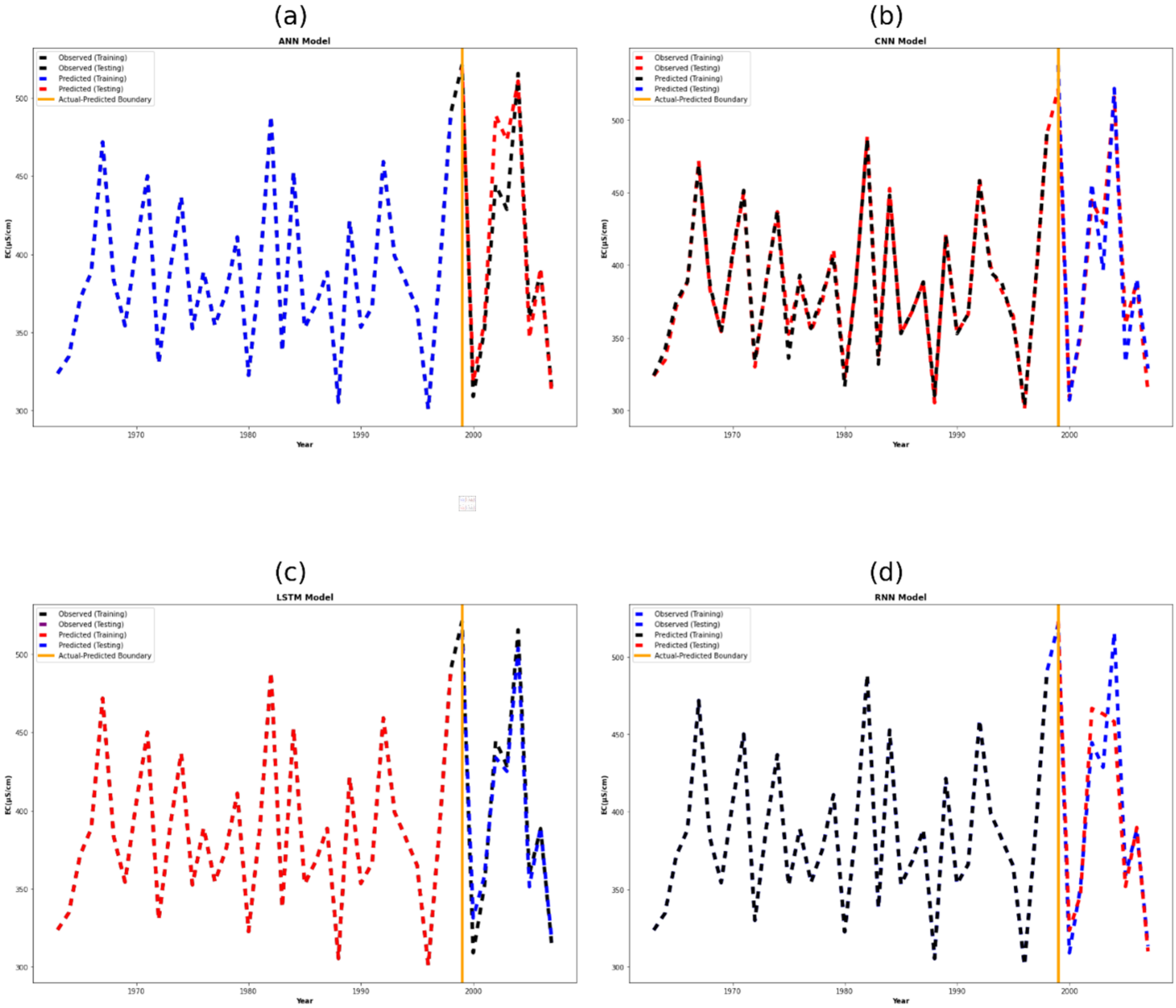

Figure 8. Training and testing of deep learning models

aid in water quality management.

The LSTM method is more sophisticated in predicting electrical conductivity compared to other machine learning methods. The LSTM method performed well during both the training (0.99) and testing (0.94) phases (33). The performance of the LSTM model is primarily attributed to its ability to capture long-range correlations in water quality data. The exceptional predictive performance of the LSTM model provides insight into the application of neural network techniques in water quality prediction (34). The (c) part of Figure 8 shows the actual vs predicted data of the LSTM model. The predicted data line is just above the actual data line in both training and testing phases.

The CNN model is placed in the last position due to its high error terms, in contrast to other machine learning models, making it unsuitable for predicting electrical conductivity. As the models are better at capturing the local patterns rather than the long-term data. Therefore, they achieved a higher R^2 of 0.99 in the training phase and 0.93 in the testing phase. However, due to higher values of other error metrics such as RMSE and MAE, it could not maintain a higher ranking. The CNN model's performance did not improve despite increasing the number of epochs. The (d) part of Figure 8 presents a comparison of actual vs predicted data of the CNN model.

Scattered plots also visualize the relationship between the actual and predicted data. The dotted points in all the scattered plots, as shown in Figure 9, compare the actual values with the predicted values. In the 1st two plots of ANN and RNN, most of the points are closely aligned with the reference line, but some of the points are still declining from the line. This shows that these models are predicting accurately but not perfectly. In the 3rd plot of the CNN model, the points exhibit a small deviation from the reference line, indicating its limited ability to predict. In the last plot of the LSTM model, all the points are closely following the reference line, indicating a strong ability to capture the time series.

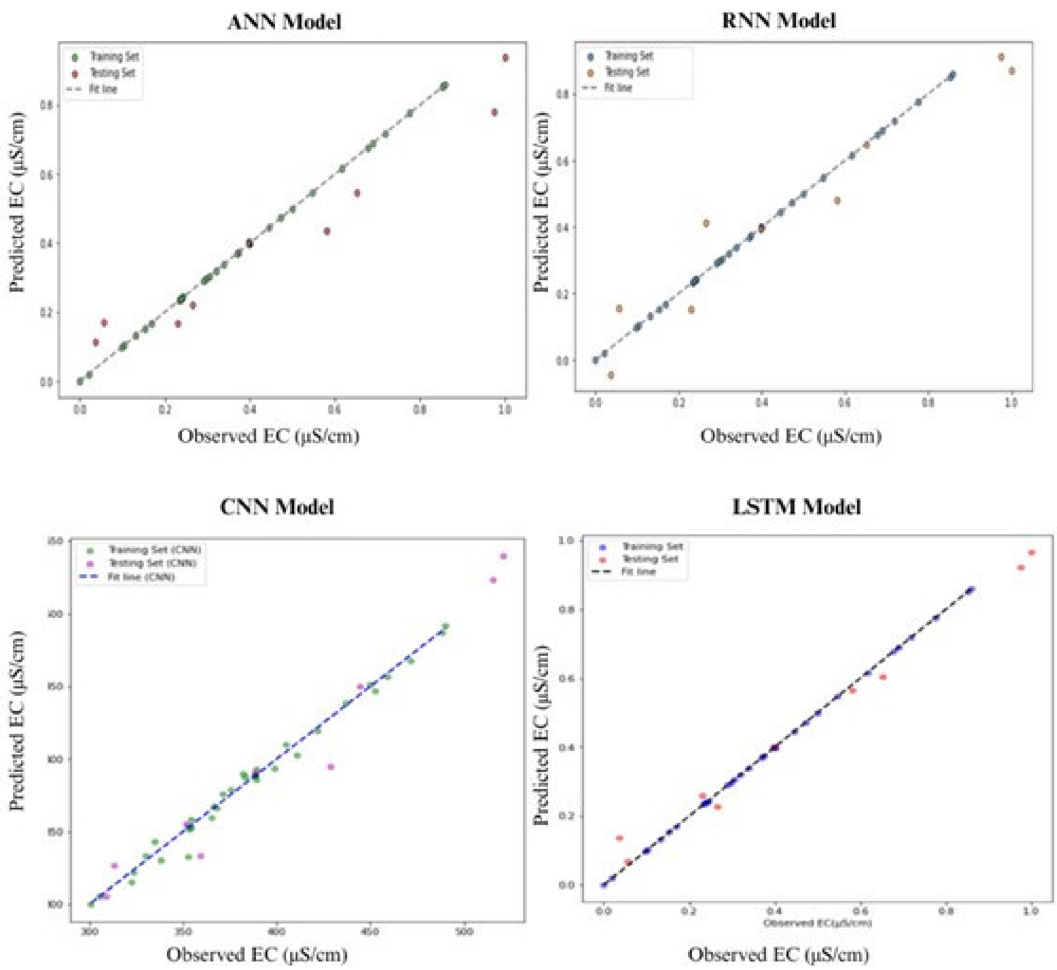

Figure 9. Observed vs predicted data of deep learning models

Models ranking via compromise programming

The statistical models cannot be ranked based on a single statistical indicator, because some models have higher R^2 values in both the training and testing phases, which is evidence of better performance; however, the same model may have higher other error metrics, leading to weaker performance. Therefore, the statistical performance indicators were merged to rank ANN, RNN, LSTM, and CNN models via compromise programming (35). The LSTM model ranked first as it excels other competing models in predicting electrical conductivity (36). The RNN, ANN, and CNN models ranked 2nd, 3rd, and 4th, respectively. The ranking of these techniques is represented by Table S1. The training, testing, and regression plots of these top-ranked models are shown in Figure 9 respectively. This comprehensive analysis, guided by compromise programming, underscores the importance of considering a broad range of indicators and scenarios when assessing the feasibility of neural network models for predicting electrical conductivity.

Discussion

The study focused on predicting the electric conductivity of water at River Bara, Khyber Pakhtunkhwa (KP), Pakistan, using advanced machine learning models. Water quality parameters are used as input and processed through machine learning models. The results were obtained in the form of statistical indicators, including R^2, RMSE, RRMSE, PBias, and NSE, as shown in Table 1. As a single statistical indicator cannot describe the performance of a statistical model, a comprehensive description requires an assessment of the overall performance of the model. Therefore, CP is an ideal technique for evaluating a model based on multiple conflicting criteria. Here, the overall performance of these models was assessed by using compromise programming.

The results demonstrate that all the models performed well, showing their potential ability for real-world applications. The LSTM model performed well in comparison to other techniques, as evident from its highest R^2 value of 0.99 during training and 0.94 during the testing phase (37). The superiority of the LSTM model is attributed to its ability to capture the complex patterns and time series dependencies in water quality data (38). Additionally, the LSTM model has its unique architecture with memory cells and a gated mechanism, which enabled it to capture the data pattern effectively and to avoid the overfitting issues.

These properties of the LSTM model make it more suitable for predicting electric conductivity. However, the other models, such as ANN, RNN, and CNN, were also better in the training phase and learned the complex nature of data easily. But at the testing stage, they failed due to the lack of an advanced memory mechanism and other problems like the vanishing gradient problem, instability of handling temporal dependencies, and sensitivity to noise. The other models were better in the training but failed to maintain their superior position in the testing phase, even though at higher epochs, due to their lower capturing capability than LSTM. The LSTM model captured the data pattern more precisely than other models at low epochs. A comprehensive comparison of these machine learning models electric on electric conductivity prediction clearly highlighted the cutting-edge ability of the LSTM model in handling complex and non-linear data relationships. A key strength of this research is to incorporate multiple input parameters, including PH, Ca, Mg, Na, PPM, Electric Conductivity, HCO_3, Temperature, Cl, Co_3, and SOR. This approach provides a historical understanding of water quality parameters.

The findings of the study perfectly align with previous studies that have consistently shown the outperformance of the LSTM model over other models in environmental forecasting. In 2023, Wu et al conducted a study on water quality prediction by Autoregressive (AR) and LSTM models and proposed that LSTM has better performance than the AR model in predicting long-term data (39). Similarly, Abbas et al utilized simple LSTM and HRU-based LSTM models for surface and sub-surface flow and concluded better results from the simple LSTM model (40). Dtissibe et al forecasted floods in the Far-North region of Cameroon in 2024 by Machine Learning and Deep Learning Models and found LSTM models as better models than the others in flood forecasting (41). Nayan et al predicted water quality using LSTM model and achieved better results (42). A study conducted by Li et al in 2023 on water quality in the Haihe River basin showed that the LSTM is an efficient model in water quality prediction (43).

Implications for water quality management

The study is conducted for River Bara at Jhansipost Station, but it applies to all the water bodies and regions having the same hydrological and environmental conditions. The adopted methodology, including statistical methods such as ANN, RNN, LSTM, and CNN with water quality parameters, is more flexible and versatile. These water quality parameters are universal, making the approaches applicable to worldwide water bodies. However, the ranking order of these models can be changed with different regions because it also depends on the data pattern. Furthermore, the research has important implications for water quality management. The findings suggest that water quality modeling can be improved using neural network models, specifically the LSTM technique, which has demonstrated excellent performance in predicting electrical conductivity. The LSTM models can predict pollution levels, enabling the regulators to implement stricter controls during high-risk periods. It can also support an early warning system for rivers to control pollution. Furthermore, the model can predict future water quality precisely based on climate change. Precise water quality prediction can help in developing proactive management strategies. Water quality prediction can help us in the formulation of water policy.

LSTM is the best model in predicting long-term time series data, but it still faces some challenges in capturing noise in small datasets with many trainable parameters. Its effectiveness is strongly dependent on the dataset size. The insufficient datasets can lead to overfitting. Moreover, due to its sequential nature and intricate architecture, it has significant computational complexity, leading to higher training time.

Conclusion

The study investigated four neural networks, ANN, RNN, LSTM, and CNN, for predicting the electric conductivity of surface waters at Bara River, Pakistan. The performance of these techniques was assessed using statistical indicators R^2, RMSE, RRMSE, MAE, NSE, and PBIAS. The LSTM model outperformed other techniques in predicting electrical conductivity based on the highest R^2 value and negligible RMSE value during the training and testing phases. This study has many practical implications for water quality management, such as proactive water quality management and well-informed decision-making. Following that, research should focus on ways to improve these models even further by experimenting with different features and hyperparameter combinations. The generalizability of the models could be improved even further by expanding the dataset to include a wider range of environmental and geographical variables.

Authors' Contributions

Conceptualization: Afed Ullah Khan, Muhammad Waqas, and Ateeq Ur Rauf.
Data curation: Afed Ullah Khan and Ateeq Ur Rauf.
Formal analysis: Muhammad Waqas, Afed Ullah Khan, and Ateeq Ur Rauf.
Funding acquisition: Muhammad Waqas.
Investigation: Fayaz Ahmad Khan, Jahanzeb Khan, and Ateeq Ur Rauf.
Methodology: Muhammad Waqas, Fayaz Ahmad Khan, and Ateeq Ur Rauf.
Project administration: Muhammad Waqas and Afed Ullah Khan.

Resources: Muhammad Waqas and Afed Ullah Khan.
Software: Muhammad Waqas.
Supervision: Afed Ullah Khan, Fayaz Ahmad Khan, and Ateeq Ur Rauf.
Validation: Fayaz Ahmad Khan, Jahanzeb Khan, and Afed Ullah Khan.
Visualization: Jahanzeb Khan and Muhammad Waqas.
Writing–original draft: Muhammad Waqas and Afed Ullah Khan.
Writing–review & editing: Afed Ullah Khan, Fayaz Ahmad Khan, and Jahanzeb Khan.

Competing interests

The authors declare that they have no competing interests. The research work is new and has not been submitted for publication elsewhere.

Ethical issues

This study did not involve any human participants or animal subjects. All data used are publicly available and were collected in compliance with relevant institutional and national research ethics guidelines.

Funding

The research work has no funding source.

Supplementary files

Supplementary file 1 contains Tables S1.

References

1. Aziz A, Akram K, Abrar ul Haq M, Hawaldar IT, Rabbani MR. Examining the role of clean drinking water plants in mitigating drinking water-induced morbidity. Sustainability. 2022;14(15):9644. doi: 10.3390/su14159644.
2. Nabi G, Ali M, Khan S, Kumar S. The crisis of water shortage and pollution in Pakistan: risk to public health, biodiversity, and ecosystem. Environ Sci Pollut Res Int. 2019;26(11):10443-5. doi: 10.1007/s11356-019-04483-w.
3. Sattari MT, Joudi AR, Kusiak A. Estimation of water quality parameters with data-driven model. J Am Water Works Assoc. 2016;108(4):E232-9. doi: 10.5942/jawwa.2016.108.0012.
4. Vats S, Sagar BB, Singh K, Ahmadian A, Pansera BA. Performance evaluation of an independent time optimized infrastructure for big data analytics that maintains symmetry. Symmetry. 2020;12(8):1274. doi: 10.3390/sym12081274.
5. Pakdaman M, Falamarzi Y, Sadoghi Yazdi H, Ahmadian A, Salahshour S, Ferrara M. A kernel least mean square algorithm for fuzzy differential equations and its application in earth's energy balance model and climate. Alex Eng J. 2020;59(4):2803-10. doi: 10.1016/j.aej.2020.06.016.
6. Mosavi A, Sajedi Hosseini F, Choubin B, Goodarzi M, Dineva AA. Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. IEEE Access. 2020;8:145564-76. doi: 10.1109/access.2020.3014908.
7. Molekoa MD, Avtar R, Kumar P, Minh HV, Kurniawan TA. Hydrogeochemical assessment of groundwater quality of Mokopane area, Limpopo, South Africa using statistical approach. Water. 2019;11(9):1891. doi: 10.3390/w11091891.
8. Madrid Y, Zayas ZP. Water sampling: traditional methods and new approaches in water sampling strategy. Trends Analyt Chem. 2007;26(4):293-9. doi: 10.1016/j.trac.2007.01.002.
9. Nabulsi R, Al-Abbadi MA. Review of the impact of water quality on reliable laboratory testing and correlation with purification techniques. Lab Med. 2014;45(4):e159-65. doi: 10.1309/lmlxnd0wnrjj6u7x.
10. Sani SA, Ibrahim A, Musa AA, Dahiru M, Baballe MA. Drawbacks of traditional environmental monitoring systems. Comput Inf Sci. 2023;16(3):30-5. doi: 10.5539/cis.v16n3p30.
11. Iqbal MM, Shoaib M, Agwanda P, Lee JL. Modeling approach for water-quality management to control pollution concentration: a case study of Ravi river, Punjab, Pakistan. Water. 2018;10(8):1068. doi: 10.3390/w10081068.
12. Khan M, Khan S, Ullah Khan A, Noman M, Usama M, Ahmad Khan F, et al. Effect of land use change on climate elasticity of water quality at multiple spatial scales. Water Pract Technol. 2022;17(11):2334-50. doi: 10.2166/wpt.2022.131.
13. Shah HA, Sheraz M, Ullah Khan A, Ahmad Khan F, Shah LA, Khan J, et al. Surface and groundwater pollution: the invisible, creeping threat to human health. Civ Environ Eng. 2020;16(1):157-69. doi: 10.2478/cee-2020-0016.
14. Zafar MU, Ahmad W. Water quality assessment and apportionment of northern Pakistan using multivariate statistical techniques–a case study. Int J Hydrol. 2018;2(1):1-7. doi: 10.15406/ijh.2018.02.00040.
15. Begum S, Firdous S, Naeem Z, Chaudhry GE, Arshad S, Abid F, et al. Combined multivariate statistical techniques and water quality index (WQI) to evaluate spatial variation in water quality. Trop Life Sci Res. 2023;34(3):129-49. doi: 10.21315/tlsr2023.34.3.7.
16. Schreiber SG, Schreiber S, Tanna RN, Roberts DR, Arciszewski TJ. Statistical tools for water quality assessment and monitoring in river ecosystems–a scoping review and recommendations for data analysis. Water Qual Res J. 2022;57(1):40-57. doi: 10.2166/wqrj.2022.028.
17. Aldrees A, Javed MF, Bakheit Taha AT, Mohamed AM, Jasiński M, Gono M. Evolutionary and ensemble machine learning predictive models for evaluation of water quality. J Hydrol Reg Stud. 2023;46:101331. doi: 10.1016/j.ejrh.2023.101331.
18. Shah MI, Javed MF, Abunama T. Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. Environ Sci Pollut Res Int. 2021;28(11):13202-20. doi: 10.1007/s11356-020-11490-9.
19. Alqahtani A, Shah MI, Aldrees A, Javed MF. Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality. Sustainability. 2022;14(3):1183. doi: 10.3390/su14031183.
20. Aslam B, Maqsoom A, Cheema AH, Ullah F, Alharbi A, Imran M. Water Quality management using hybrid machine learning and data mining algorithms: an indexing approach. IEEE Access. 2022;10:119692-705. doi: 10.1109/access.2022.3221430.
21. Shah MI, Alaloul WS, Alqahtani A, Aldrees A, Musarat MA, Javed MF. Predictive modeling approach for surface

water quality: development and comparison of machine learning models. Sustainability. 2021;13(14):7515. doi: 10.3390/su13147515.
22. Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient water quality prediction using supervised machine learning. Water. 2019;11(11):2210. doi: 10.3390/w11112210.
23. Rehman SS, Sabir MA, Khan J. Discharge characteristics and suspended load from rivers of Northern Indus Basin, Pakistan. J Himal Earth Sci. 1997;30(1):325-36.
24. Mehmood A, Jia S, Lv A, Zhu W, Mahmood R, Saifullah M, et al. Detection of spatial shift in flood regime of the Kabul river basin in Pakistan, causes, challenges, and opportunities. Water. 2021;13(9):1276. doi: 10.3390/w13091276.
25. Patro SG, Sahu KK. Normalization: A Preprocessing Stage. ArXiv [Preprint]. March 19, 2015 [Cited 2025 January 14]. Available from: https://arxiv.org/abs/1503.06462.
26. Ullah B, Fawad M, Ullah Khan A, Mohamand SK, Khan M, Iqbal MJ, et al. Futuristic streamflow prediction based on CMIP6 scenarios using machine learning models. Water Resour Manag. 2023;37(15):6089-106. doi: 10.1007/s11269-023-03645-3.
27. Iqbal Z, Shahid S, Ahmed K, Ismail T, Ziarh GF, Chung E-S, et al. Evaluation of CMIP6 GCM rainfall in mainland Southeast Asia. Atmos Res. 2021;254:105525. doi: 10.1016/j.atmosres.2021.105525.
28. Shiru MS, Chung E-S. Performance evaluation of CMIP6 global climate models for selecting models for climate projection over Nigeria. Theor Appl Climatol. 2021;146(1):599-615. doi: 10.1007/s00704-021-03746-2.
29. Singh G, Chaudhary S, Gupta D, Kumar Mishra V. Assessing the water quality of River Ganga in Varanasi, India, through WQI, NPI, and multivariate techniques: a comprehensive study. Water Pract Technol. 2024;19(4):1099-118. doi: 10.2166/wpt.2024.027.
30. Khan M, Ullah Khan A, Khan S, Ahmad Khan F. Assessing the impacts of climate change on streamflow dynamics: a machine learning perspective. Water Sci Technol. 2023;88(9):2309-31. doi: 10.2166/wst.2023.340.
31. Rustam F, Ishaq A, Kokab ST, de la Torre Diez I, Mazón JL, Rodríguez CL, et al. An artificial neural network model for water quality and water consumption prediction. Water. 2022;14(21):3359. doi: 10.3390/w14213359.
32. Aslan S, Zennaro F, Furlan E, Critto A. Recurrent neural networks for water quality assessment in complex coastal lagoon environments: a case study on the Venice Lagoon. Environ Model Softw. 2022;154:105403. doi: 10.1016/j.envsoft.2022.105403.
33. Liu P, Wang J, Sangaiah AK, Xie Y, Yin X. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability. 2019;11(7):2058. doi: 10.3390/su11072058.
34. Hu Z, Zhang Y, Zhao Y, Xie M, Zhong J, Tu Z, et al. A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors (Basel). 2019;19(6):1420. doi: 10.3390/s19061420.
35. Khan S, Ullah Khan A, Khan M, Ahmad Khan F, Khan S, Khan J. Intercomparison of SWAT and ANN techniques in simulating streamflows in the Astore Basin of the Upper Indus. Water Sci Technol. 2023;88(7):1847-62. doi: 10.2166/wst.2023.299.
36. Zhou J, Wang Y, Xiao F, Wang Y, Sun L. Water quality prediction method based on IGRA and LSTM. Water. 2018;10(9):1148. doi: 10.3390/w10091148.
37. Chandra R, Goyal S, Gupta R. Evaluation of deep learning models for multi-step ahead time series prediction. IEEE Access. 2021;9:83105-23. doi: 10.1109/access.2021.3085085.
38. Liu P, Wang J, Sangaiah AK, Xie Y, Yin X. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability. 2019;11(7):2058. doi: 10.3390/su11072058.
39. Wu X. Water quality prediction based on AR and LSTM model. J Phys Conf Ser. 2023;2580(1):012019. doi: 10.1088/1742-6596/2580/1/012019.
40. Abbas A, Baek S, Kim M, Ligaray M, Ribolzi O, Silvera N, et al. Application of Deep Recurrent Neural Networks for Modeling Surface and Sub-Surface Flow at High Temporal Resolution. EGU General Assembly; 2020. p. 6216. doi: 10.5194/egusphere-egu2020-6216.
41. Dtissibe FY, Abba Ari AA, Abboubakar H, Njoya AN, Mohamadou A, Thiare O. A comparative study of machine learning and deep learning methods for flood forecasting in the Far-North region, Cameroon. Sci Afr. 2024;23:e02053. doi: 10.1016/j.sciaf.2023.e02053.
42. Nayan AA, Khan MS, Ferdaous J, Mozumder AN, Alam MK, Kibria MG. Drinking water quality analysis and prediction using LSTM: safe drinking water for school children. In: Arefin MS, Kaiser MS, Bhuiyan T, Dey N, Mahmud M, eds. Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning. Vol 867. Singapore: Springer, Singapore; 2024. p. 499-513. doi: 10.1007/978-981-99-8937-9_34.
43. Li Q, Yang Y, Yang L, Wang Y. Comparative analysis of water quality prediction performance based on LSTM in the Haihe river basin, China. Environ Sci Pollut Res Int. 2023;30(3):7498-509. doi: 10.1007/s11356-022-22758-7.