Journal of Water Process Engineering xxx (xxxx) 106595

Contents lists available at ScienceDirect

Journal of Water Process Engineering

journal homepage: www.elsevier.com/locate/jwpe

A novel hybrid deep learning model for real-time monitoring of water pollution using sensor data

Majid Bagheri [a,*], Karim Bagheri [b], Nakisa Farshforoush [c], Antonio Velazquez [a], Ying Liu [a]

[a] Department of Engineering Technology, Savannah State University, Savannah, GA, USA
[b] Department of Chemical Engineering, Ilam University, Ilam, Iran
[c] Department of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran

ARTICLE INFO

Editor: Soroush Abolfathi

Keywords:
Hybrid model
CNN-LSTM
Dissolved oxygen
XGBoost
Significant predictors

ABSTRACT

This study developed a hybrid model for predicting dissolved oxygen (DO) using real-time sensor data for thirteen parameters. This novel hybrid model integrated one-dimensional convolutional neural networks (CNN) and long short-term memory (LSTM) to improve the accuracy of prediction for DO in water. The hybrid CNN-LSTM model predicted DO concentration in water using soft sensor data. The primary input parameters to the model were temperature, pH, specific conductivity, salinity, density, chlorophyll, and blue-green algae. The model used 38,681 water quality data for training and testing the hybrid deep learning network. The training procedure for the model was successful. The training and test losses were both nearly zero and within a similar range. With a coefficient of determination (R^2) of 0.94 and a mean squared error (MSE) of 0.12, the hybrid model indicated higher performance compared to the classical models. The normal distribution of residual errors confirmed the reliability of the DO predictions by the hybrid CNN-LSTM model. Feature importance analysis indicated pH as the most significant predictor and temperature as the second important predictor. The feature importance scores based on extreme gradient boosting (XGBoost) for the pH and temperature were 0.76 and 0.12, respectively. This study indicated that the hybrid model can outperform the classical machine learning models in the real-time prediction of DO concentration.

Nomenclature

DO	dissolved oxygen
BOD	biochemical oxygen demand
AI	artificial intelligence
ML	machine learning
CNN	convolutional neural networks
DNN	deep neural networks
RNN	recurrent neural networks
LSTM	long short-term memory
AUV	autonomous underwater vehicle
Min	minimum
Max	maximum
STD	standard deviation
IQR	interquartile range
t-SNE	t-distributed stochastic neighbor embedding
XGBoost	extreme gradient boosting
MSE	mean squared error
RMSE	root mean squared error
R^2	coefficient of determination
relu	rectified linear unit
tanh	hyperbolic tangent function
sigmoid	sigmoid function
adam	adaptive moment optimizer
sgd	stochastic gradient descent
rmsprop	root mean squared propagation
RF	random forest
SVR	support vector regression
MLP	multilayer perceptron
MAE	mean absolute error

* Corresponding author at: Department of Engineering Technology, Savannah State University, Savannah, GA 31404, USA.
E-mail address: bagherim@savannahstate.edu (M. Bagheri).

https://doi.org/10.1016/j.jwpe.2024.106595
Received 5 August 2024; Received in revised form 15 October 2024; Accepted 16 November 2024
2214-7144/

1. Introduction

Emerging and fugitive contaminants, such as pharmaceuticals, personal care products, pesticides, and microplastics, have polluted aquatic and freshwater systems due to their widespread use and inadequate treatment processes [1]. These emerging and fugitive contaminants persist in the environment and bioaccumulate in aquatic organisms. Microplastics can impact water quality by serving as both physical contaminants and transporters of toxic chemicals, such as heavy metals and pesticides [2]. These emerging and fugitive contaminants can have long-term, cumulative effects on both the environment and human health. Thus, novel technologies, including advanced monitoring tools, are an essential part of water management systems to effectively control these contaminants.

Measuring important parameters such as dissolved oxygen (DO) and biochemical oxygen demand (BOD), which influence the quality of water, is a challenging task in water quality monitoring [3]. Despite the availability of low-cost sensors for water quality measurements, some water characteristics still require a laboratory approach for analysis due to the lack of online real-time sensors [4]. Laboratory-based methods for measuring important water quality parameters encounter challenges such as time-consuming sample collection processes, delays in obtaining results, costly procedures, and variability in sampling techniques and storage conditions [5]. Addressing these challenges requires efforts to improve sample collection techniques, enhance laboratory capacity, and integrate complementary monitoring approaches to provide comprehensive water quality assessments [6]. Soft sensor-based monitoring is a reliable approach to address the challenges of water quality monitoring when a specific physical measurement sensor is missing. A soft sensor is a virtual sensing method that creates an inferential model to estimate different parameters of interest based on the data for other available measured parameters [7–9].

Artificial intelligence (AI) and machine learning (ML) are highly useful for water quality modeling due to their ability to process vast amounts of data from various sources, such as sensors, satellite imagery, and environmental databases [10–12]. The AI and ML models can accurately predict water quality parameters by considering complex relationships between environmental factors, such as temperature, pH levels, and pollutant sources [13–15]. AI and ML models improve over time when they are fed with more data, especially from real-time in-situ physical sensors [16]. This makes the models valuable tools for monitoring and managing water resources, facilitates early detection of pollution events, and guides decision-making for water quality management efforts [17]. The results of numerous studies confirm the higher predictive performance of the AI and ML models compared to the traditional models for water quality monitoring [18].

Deep learning is a branch of machine learning that has received more attention for water quality monitoring in recent years [19,20]. Deep learning models have outperformed classical models in predicting water quality parameters due to their ability to handle high-dimensional data. These models can capture intricate relationships between various environmental factors and water quality indicators [21]. Deep learning models, such as convolutional neural networks (CNN), deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) can effectively process diverse datasets [22–24]. The models can optimize their predictions over time and enhance their accuracy and reliability in forecasting water quality parameters such as DO levels and pollutant concentrations [25]. Thus, deep learning models have fundamental advantages for real-time water monitoring and early warning systems that contribute to more effective resource conservation and pollution control efforts [26].

Hybrid deep learning models combine the strengths of multiple deep learning algorithms to improve the accuracy of predictions [27–29]. A CNN-LSTM model is superior to regular deep learning models for predicting DO concentration because it combines the strengths of both CNNs and LSTM networks. This allows the CNN-LSTM model to effectively handle both spatial and temporal aspects of water quality data. In this ensemble learning method, each deep learning model is trained independently, and the outputs of the models are combined to make predictions [30]. This approach helps in reducing overfitting and improving generalization performance, thus enhancing predictive accuracy [31]. Such a hybrid model, with its improved predictive accuracy, can play a major role in the real-time monitoring of water quality parameters using real-time data recorded by in-situ sensors. The hybrid CNN-LSTM model is useful for real-time prediction of water quality parameters when the data for a specific sensor is not available.

This research develops a hybrid one-dimensional CNN-LSTM model to predict DO concentration in water based on real-time sensor data for thirteen other water quality parameters. The hybrid CNN-LSTM is a robust model to predict DO concentration. It uniquely combines CNNs for extracting spatial features and LSTMs for capturing temporal dependencies. The hybrid model offers a more comprehensive approach for water quality data. The primary inputs to the model are the sensor locations and water quality parameters, including temperature, pH, specific conductivity, salinity, density, chlorophyll, and blue-green algae. The hybrid CNN-LSTM forms by merging the output layers of the CNN and LSTM. This novel hybrid model enhances the accuracy of predictions for DO concentration in water compared to the classical ML models.

2. Methodology

2.1. Study area and water quality dataset

The Savannah River is a major river in the southeast United States. The river flows along the border between Georgia and South Carolina. It stretches about 484 km, beginning in the Blue Ridge Mountains of North Carolina, where the Tugaloo and Seneca rivers converge to form the Savannah. The river flows southeast through the Piedmont and Coastal Plain regions, eventually emptying into the Atlantic Ocean near the city of Savannah, Georgia. The original data for the study area were measured during six days of spatial water quality survey collection. The study area was close to the Georgia Power Plant McIntosh in Effingham County, Georgia, and the Hutchinson Island DO injector site. The water quality data included measurements at multiple locations and depths using physical sensors [32].

An IVER3 EcoMapper Autonomous Underwater Vehicle (AUV; SN 3086) measured the water quality characteristics. This instrument measures water quality data using a YSI EXO1 Water Quality Sonde. An onboard depth sounder and pressure transducer measured the AUV bathymetric data. The calibration of the AUV and EXO2 Sonde before and after each data collection trip was based on the water quality sampling protocols [33]. Water quality data measurements were generally at or near slack-tide conditions at the surface or 2 ft below the surface. YSI sensors 599870-01, 599870-01, 599102-0, 599102-01, 599100-01, 57760, and 599101-01 measured specific conductivity, temperature, total algae, chlorophyll, DO, pH, and turbidity. To develop the hybrid CNN-LSTM model in this study, the dataset included 38,681 measurements of water quality characteristics for the study area. Table 1 represents a summary of statistics for the data used to develop the hybrid CNN-LSTM model. To predict the DO concentration in water, the CNN-LSTM model used thirteen water quality parameters as inputs.

2.2. Exploratory data analysis

To detect hidden patterns in the water quality dataset, a preliminary data analysis was performed using t-distributed stochastic neighbor embedding (t-SNE). This method involves dimensionality reduction to reveal and visualize patterns within complex datasets [34]. This approach focuses on preserving local relationships and capturing the intrinsic structure of the data. The method begins by computing pairwise simi-

Table 1
Summary of statistics for the input and output parameters used in the modeling processes.

Water quality (unit)	Min.	Max.	Mean	Median	Std.	IQR
Inputs parameters:						
Pitch angle	−33.7	18.9	−0.8	−2.3	4.11	6.5
Roll angle	−28.4	23.5	1.6	1.3	2.52	1.8
Depth to surface (ft)	0	6.3	1.6	2.1	0.88	1.9
Depth to bottom (ft)	5	32.1	16.3	16.5	5.12	7.6
Total water column (ft)	5.2	33.9	17.9	18.2	5.66	8.6
Current step	1	44	22.4	22	12.5	23
Temperature (C)	28.5	30.3	28.9	28.8	0.39	0.5
Specific conductivity (uS/cm)	141	3870	487	248	621.9	130
pH	6.5	7	6.7	6.7	0.09	0.1
Chlorophyll (RFU)	1.6	3.8	2.4	2.2	0.42	0.6
Blue-Green algae (RFU)	0.1	2	0.5	0.5	0.18	0.2
Salinity (ppt)	0.1	2	0.2	0.1	0.32	0.1
Density (kg/m^3)	996	997	996	996	0.23	0
Output parameter:						
Dissolved oxygen (mg/L)	5.6	7.7	6.3	6.2	0.46	0.5

Min: Minimum, Max: Maximum, Std: Standard deviation, IQR: Interquartile range.

larities and measures the similarity between two data points based on their Euclidean distance [35]. It converts the similarities into conditional probabilities and aims to find a lower-dimensional representation where the conditional probabilities of the data points are as similar as possible to those in the original space. This exploratory data analysis method is effective for identifying clusters that might be challenging to discern in high-dimensional spaces [36]. The t-SNE technique facilitates the visualization of intricate patterns within the water quality datasets by mapping data points to a lower-dimensional space.

2.3. Feature importance analysis

The contribution of the individual input parameters to the performance of the CNN-LSTM model was examined by feature importance analysis [37]. Extreme gradient boosting (XGBoost) determined the significant variables for predicting DO concentration in water. XGBoost calculates importance scores for the input features based on how they contribute to the model's predictive performance [38,39]. The method calculates the feature importance score using the weighted improvement in the performance of the model. Reduction in the objective loss function is a criterion to measure the performance improvement. The MSE was the loss function to measure the difference between a predicted output and the actual target value. The XGBoost regression was developed using *scikit-learn* in Python [40]. The grid search cross-validation technique optimized the hyperparameters of the XGBoost. The model showed the highest predictive performance, with a maximum depth of 4 and a learning rate of 0.15. The XGBoost technique calculates the importance scores based on the total gain of each feature across all nodes and the number of trees in the ensemble. The equation for calculating feature importance scores in XGBoost is as follows:

$$\text{Feature importance} = \sum_{\text{nodes}} \text{Gain}_{\text{nodes}} / \text{Number of trees} \tag{1}$$

2.4. Hybrid CNN-LSTM model

A convolutional neural network (CNN) is a deep learning model that consists of multiple convolutional layers, pooling layers, and fully connected layers [41]. Convolutional layers apply filters to inputs in order to capture spatial hierarchies and detect important features. The model applies an activation function element-wise to the output of each convolutional operation. The activation function adds non-linearity to the network [42]. Pooling layers reduce the dimensionality of the feature maps and make computation more efficient while preserving important information [43]. Fully connected layers perform high-level reasoning and decision-making based on the extracted features. Through forward and backward propagation, CNNs are trained to automatically learn and extract relevant patterns from input data and predict the output [44]. The mathematical representation of the CNN model with a fully connected final layer is as follows:

$$z_i = \sigma_i \left(W_i \times x_{i-1} + b_i \right) \tag{2}$$

$$y = \sigma \left(W_{final} \times z_{previous} + b_{final} \right) \tag{3}$$

where z_i is the output of layer *i*, σ_i is the activation function of layer *i*, W_i is the weight of layer *i*, b_i is the bias of layer *i*, x_{i-1} is the input to layer *i*, and *y* is the predicted output.

Long short-term memory (LSTM) is a recurrent neural network that effectively works for sequential data by mitigating the vanishing gradient problem [45]. LSTM achieves this by incorporating memory cells and gating mechanisms that regulate the flow of information. Each LSTM unit contains a cell state serving as the memory and three gates controlling the flow of information (input, forget, and output gates). The input gate determines which information to update, the forget gate discards unwanted information, and the output gate regulates which information to output [46]. By dynamically adjusting these gates based on the input and the current state, LSTM networks can learn to effectively process and predict sequences, such as time series predictions. The following equations summarize the computations for the LSTM at each time step.

$$i_t = \sigma \left(W_i.x_t + U_i.h_{t-1} + b_i \right) \tag{4}$$

$$f_t = \sigma \left(W_f.x_t + U_f.h_{t-1} + b_f \right) \tag{5}$$

$$o_t = \sigma \left(W_o.x_t + U_o.h_{t-1} + b_o \right) \tag{6}$$

where i_t is the input gate, f_t is the forget gate, o_t is the output gate, x_t is the input at time step *t*, h_t is the hidden state at time step *t*, σ is the activation function, *W* and *U* are the weight matrices, and *b* is the bias vector.

The CNN-LSTM model development included separate designs for the one-dimensional CNN and LSTM architectures based on the train datasets. The CNN and LSTM models were merged by concatenating their layers to form a new layer. The hybrid CNN-LSTM model had a fully connected output layer. The inputs of this output layer were the merged layers (Fig. 1). The output layer of the hybrid model had a linear activation function to predict the DO concentration.

2.5. Model training procedure

The training procedure for the hybrid CNN-LSTM model involved several key steps. The preprocessing of the water quality dataset was the first step to appropriately adjust the data for the model. The preprocessing included dividing the water quality dataset into training and test sets and reshaping them for both the CNN and LSTM models. The two models had separate architectures with independent training processes. The networks used 80 % of the data for training and 20 % for testing the models. The hybrid model integrated CNN layers to capture spatial features and LSTM layers to model temporal dependencies. The hybrid CNN-LSTM model formed a single layer by combining the output of the CNN layers and the LSTM layers.

The CNN model was a one-dimensional convolutional neural network with 64 computational neurons in the first layer and a kernel size of 3. Grid search cross-validation determined the most optimal hyperparameters for the model (Table 2). Several activation functions were applied to train the model. The grid search cross-validation showed a rectified linear unit as the best activation function [47]. The model had a max-pooling layer with a size of 2. The second convolutional layer

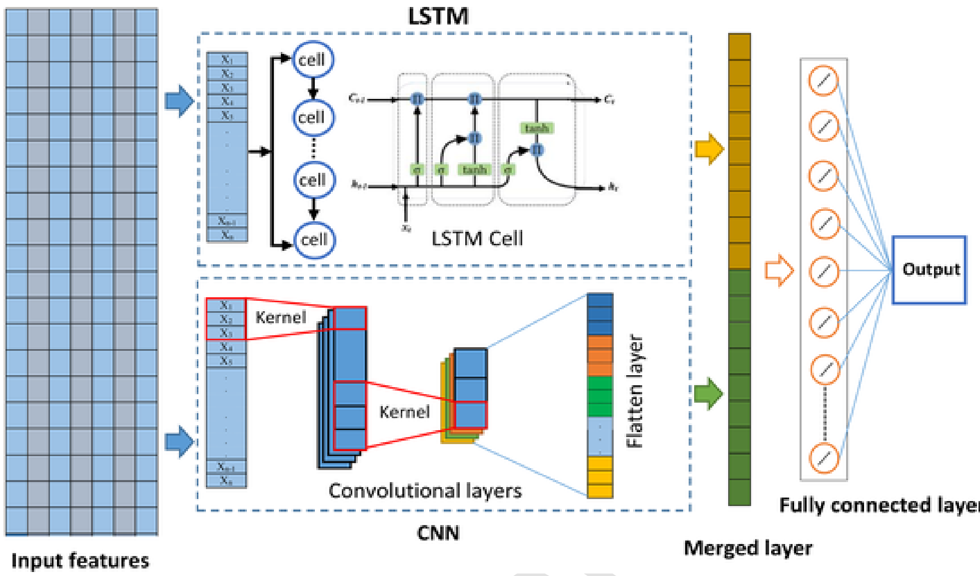

Fig. 1. Architecture of the hybrid CNN-LSTM model for predicting DO concentration.

Table 2
Grid search cross-validation for the hyperparameters of the model.

Parameter	Search space	Optimal value
Activation function	['relu', 'tanh', 'sigmoid']	'relu'
Optimizer	['adam', 'sgd', 'rmsprop']	'adam'
CNN filter size	[32, 64, 128]	64
CNN kernel size	[2,3,4]	3
LSTM units	[32, 64, 128]	64
Dropout rate	[0.2, 0.3, 0.4]	0.2
Batch size	[500, 1000, 2000]	1000
Epochs	[100, 500, 1000]	100

had 32 computational neurons and a kernel size of 3. The size of the max-pooling layer for the second convolutional layer was 2. The CNN model also had a flattening layer to prepare the inputs for the next layer. The grid search cross-validation indicated that the LSTM model has the highest performance with 64 computational cells. The output layers of the CNN and LSTM models were merged into a fully connected dense layer with 128 computational neurons to form the hybrid model. The activation function of the hybrid CNN-LSTM model was a rectified linear unit.

The model compilation included an MSE for the loss function, an adaptive moment optimizer (adam), and evaluation metrics. The training processes consisted of feeding batches of 1000 of the training data into the model. The backpropagation method computed the gradients and updated the weights accordingly. Performance of the model was monitored on the validation dataset to tune hyperparameters. The validation dataset was 20 % of the training dataset. The training of the model was repeated for 100 epochs. Regularization was used to prevent overfitting and ensure the model generalizes well to unseen data. Dropout is one of the most common regularization techniques in neural networks, including hybrid models like CNN-LSTM. The dropout rate of 0.2 was the optimal value for the hybrid model according to the grid search cross-validation. The hybrid CNN-LSTM model was developed in Python 3.11 using the TensorFlow platform. Fig. 2 demonstrates various steps of designing and developing the hybrid CNN-LSTM model.

The training procedure for the model was successful. The near values for the training and test losses over 100 computational epochs confirm the reliability of the predictions by the hybrid CNN-LSTM model (Fig. 3). When training and test losses are both in the same range and very near zero, it indicates that the hybrid model has effectively learned to generalize from the training data to new unseen data [48]. The predictive CNN-LSTM model can achieve a balance in its ability to capture underlying patterns without memorizing the noise of water quality data. Such results signify a well-optimized model that effectively learns from the water quality data without exhibiting either high bias or high variance. This is a sign of the robustness and generalization capabilities of the hybrid CNN-LSTM model [49].

2.6. Performance evaluation

Evaluating the performance of the hybrid CNN-LSTM model is a critical step in assessing the effectiveness and reliability of the DO predictions. Mean squared error (MSE) and root mean squared error (RMSE) are common loss functions in regression problems. This study used these error metrics to measure the difference between predicted values and actual values of the DO concentration [50]. The lower values of MSE, RMSE, and mean absolute error (MAE) indicate better performance [51]. The coefficient of determination (R^2) was the criteria to evaluate the goodness of fit [52,53]. The MSE, RMSE, MAE, and R^2 can be calculated by the following equations.

$$\mathrm{MSE} = 1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{pi}} - \mathrm{y}_{\mathrm{ti}}\right)^2 \tag{7}$$

$$\mathrm{RMSE} = \sqrt{1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{pi}} - \mathrm{y}_{\mathrm{ti}}\right)^2} \tag{8}$$

$$\mathrm{MAE} = 1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left|\mathrm{y}_{\mathrm{pi}} - \mathrm{y}_{\mathrm{ti}}\right| \tag{9}$$

$$\mathrm{R}^2 = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{ti}} - \mathrm{y}_{\mathrm{pi}}\right)^2 \Big/ \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{y}_{\mathrm{ti}} - \overline{\mathrm{y}}\right)^2 \tag{10}$$

where $\overline{\mathrm{y}}$ is the average of y over the *n* data, y_t is the actual value, and y_p is the predicted value.

3. Results and discussion

3.1. Principal components and patterns

The outputs of pattern recognition using t-SNE to visualize the possible clusters in the water quality dataset are shown in Fig. 4. The results indicated relatively separate clusters in the water quality data. The analysis showed separate clusters for high DO concentrations and low

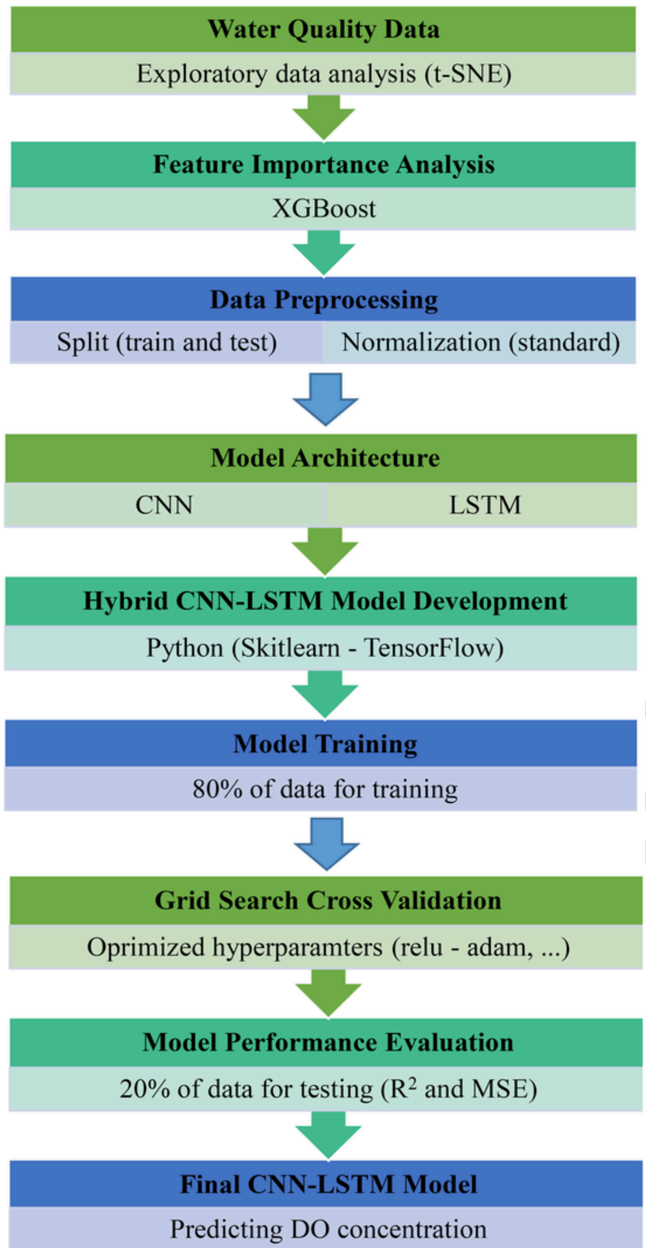

Fig. 2. Steps for designing and developing the hybrid CNN-LSTM model for predicting DO concentration in water.

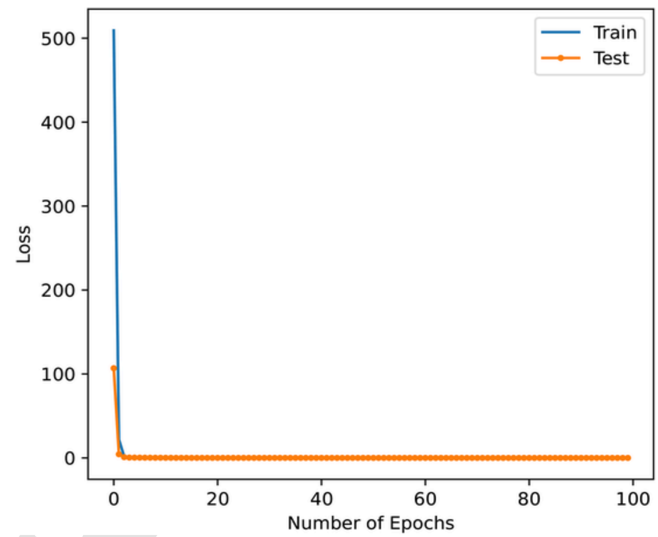

Fig. 3. Loss of the hybrid CNN-LSTM model for both training and test datasets for the hundred computational epochs.

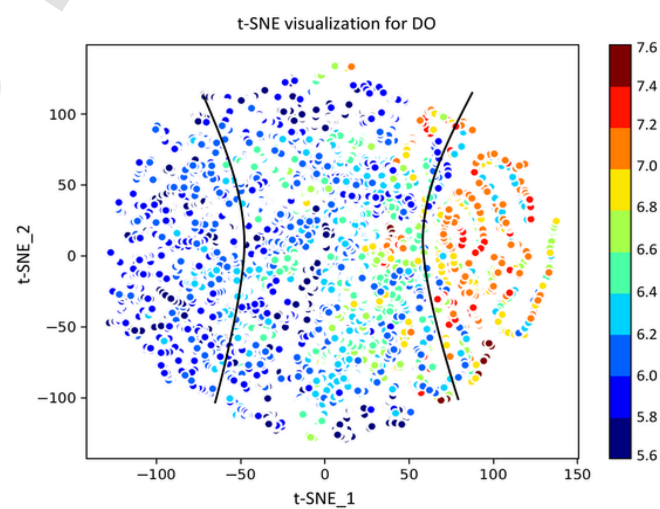

Fig. 4. Visualization of the hidden patterns in the water quality dataset using t-SNE technique.

DO concentrations. As shown in Fig. 4, the clusters were distinguishable based on the first t-SNE dimension. This indicates that some input parameters were more significant than other parameters. The possibility of different zones of water quality needs further investigation by applying t-SNE to water quality data for a longer period of time. The t-SNE, together with other clustering algorithms, is useful for deeper cluster analysis of water quality data with longer periods [54].

3.2. Significant predictors

The results of feature importance analysis using XGBoost for the thirteen input parameters were insightful. The final score for each input parameter was converted to relative importance and visualized using a bar chart (Fig. 5). With the highest relative importance score for the pH, the analysis indicated pH as the most important predictor for predicting DO concentration in water. The pH had a relative importance of 0.76, and then temperature was the second important predictor, with a relative importance of 0.12. The results of previous studies indicated that pH and temperature are important parameters for predicting DO concentration in water [55,56]. The high importance of pH and temperature can be attributed to their fundamental roles in both chemical and biological processes that govern oxygen dynamics in water. pH directly affects the chemical equilibrium of DO. It also influences photosynthesis and respiration, which are critical for oxygen production and consumption. Temperature affects oxygen consumption by influencing the solubility of oxygen in water and the metabolic rates of aquatic organisms. These combined effects of pH and temperature are more significant and consistent in predicting DO concentration compared to other variables like salinity or conductivity, which have a more indirect influence on oxygen levels [57].The analysis indicated that other input parameters such as chlorophyll, specific conductivity, and blue-green algae have a relative importance of less than 0.04. The results of the feature importance analysis are in agreement with the separate clusters based

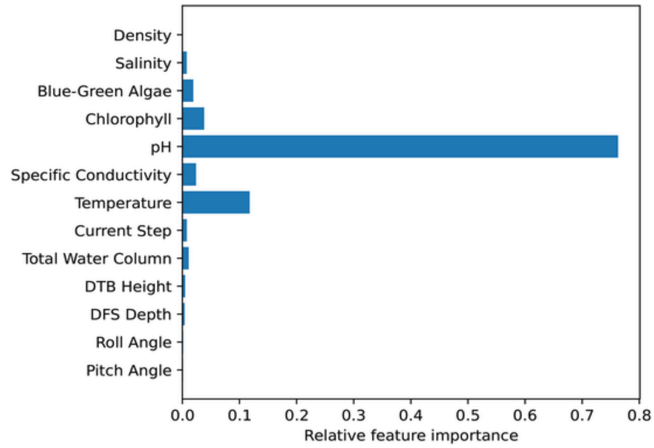

Fig. 5. The relative importance of the thirteen features for the prediction of DO concentration using the extreme gradient boosting technique.

on the first principal component of the t-SNE. The results of these analyses indicated that accurate prediction of DO concentration is possible from the data recorded by a few other physical sensors (data for pH).

3.3. *Predictive performance of the hybrid CNN-LSTM model*

The hybrid model showed improved accuracy for the prediction of DO concentration based on the test dataset for water quality. The hybrid CNN-LSTM model captured the picks of water quality data for DO concentration (Fig. 6). The results proved the high generalization capability of this predictive model for detecting water quality trends [58]. This high generalization enables the hybrid CNN-LSTM model to identify potential contamination sources, predict water quality fluctuations, and recommend proactive measures to maintain or improve water quality standards. The results of the modeling are particularly important for in-situ real-time monitoring of water pollution using tiny machine learning systems. The microcontrollers equipped with the proposed hybrid model can autonomously and effectively monitor water pollution occurrences. As shown in Fig. 6, the predictive values for DO concentration match the measured values relatively well for both training and test datasets. Goodness of fit indicates how closely the predicted values match the actual values in the dataset used for training and validation of the model [59]. A reliable predictive model typically exhibits a high goodness of fit, meaning that it accurately captures the patterns and relationships within the water quality data.

The results also indicated that the hybrid CNN-LSTM model has high performance based on the R^2 and the error metrics (Fig. 7). The model achieved the highest accuracy with an R^2 of 0.94 and an MSE of 0.12 for the test dataset. The RMSE for the test dataset was 0.34. The hybrid CNN-LSTM model indicated higher predictive performance compared to the classical machine learning models. The previous studies indicated that fuzzy neural networks predict DO concentration in water with accuracies up to 92 % [60]. In this study, hourly DO and water quality variables for a one-year period were used in the modeling process. The MAE for the attention-based recurrent neural networks varied from 0.16 to 0.18 [61]. Table 3 presents the comparison between the predictive performance of the proposed CNN-LSTM model and other classical models in previous studies. The hybrid CNN-LSTM model outperforms single models in predicting water quality due to its ability to capture complex relationships and patterns inherent in diverse water quality data [62–67]. By combining two powerful deep learning algorithms, the hybrid model can effectively handle the multidimensional nature of water quality parameters [68]. The model offers enhanced flexibility in representing both linear and nonlinear relationships. The model utilizes ensemble learning techniques to reduce variance and improve prediction accuracy. It leverages transfer learning to adapt pre-trained representations for water quality prediction tasks.

The residual errors of predicting DO by the hybrid CNN-LSTM model based on the training and test datasets were plotted against the frequency in Fig. 8. The results for residual errors demonstrated a normal distribution. A normal distribution (Gaussian curve) shows the highest error in the middle of the curve [69]. Based on the test data, the Gaussian curve indicated that the DO predictions were symmetrical.

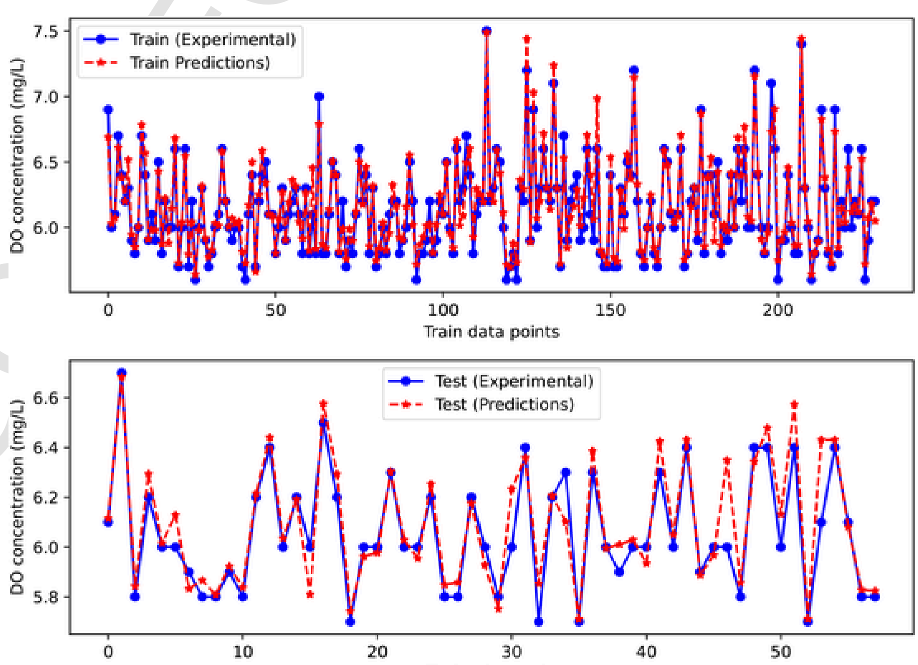

Fig. 6. Predictions of the DO concentration by the hybrid CNN-LSTM model vs. target values based on training and test datasets.

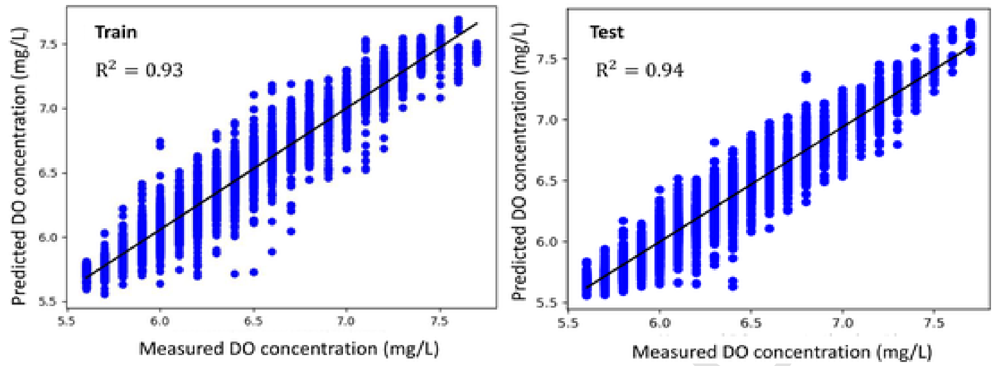

Fig. 7. Regression analysis for the hybrid CNN-LSTM model for predicting DO concentration: measured DO values vs. predicted DO values.

Table 3
Predictive performance of the hybrid CNN-LSTM model compared to the reported studies.

Model	Input variables	R^2	Error	Reference
CNN-LSTM	Thirteen water quality parameters and sensor data	0.94	MSE = 0.12 RMSE = 0.34	This study
RF	Water quality parameters	0.67	MAE = 0.89 RMSE = 1.28	Garabaghi et al. [62]
SVR	Temperature and flowrate	0.96	MAE = 0.57 RMSE = 0.64	Dehghani et al. [63]
MLP	Temperature, pH, specific conductance, and chemical oxygen demand	–	MAE = 0.49 RMSE = 0.65	Ahmed and Lin [64]
LSTM	seven hydrometeorological variables	0.90	RMSE = 0.37	Zhi et al. [65]
RNN	Temperature, specific conductance, streamflow discharge, pH	–	MAE = 0.25 RMSE = 0.43	Moghadam et al. [66]
LSTM-RNN	Temperature and DO concentration	0.95	MSE = 0.54 MAE = 0.42	Pan et al. [67]

RF: Random forest, **SVR:** Support vector regression, **MLP:** Multilayer perceptron, **MAE:** Mean absolute error.

The frequency of error bunched in the middle and died off in the tails of the curve. The results demonstrated that the error distribution for predicting DO concentration in water was around zero. The residual errors confirmed that the hybrid CNN-LSTM model can accurately and reliably predict DO concentration in water.

The predictive hybrid CNN-LSTM model is potent for practical implementation. The model can be integrated into existing water quality monitoring systems that utilize sensor networks. In such setups, multiple sensors deployed in different locations of rivers continuously collect data on water quality parameters. By feeding this hybrid CNN-LSTM model with real-time data from the sensor networks, water quality monitoring can become more predictive rather than reactive. The hybrid CNN-LSTM model can provide early warnings about potential drops in DO concentration. It also allows environmental authorities to respond proactively to prevent harmful algal blooms. The system can support informed decision-making in water resource management, such as optimizing aeration systems in reservoirs and adjusting wastewater discharge regulations based on forecasted DO levels.

4. Conclusions

This study introduced a novel hybrid CNN-LSTM model to accurately predict DO concentration in water. This study demonstrated the efficacy of this hybrid deep learning model in accurately predicting DO concentration using real-time sensor data compared to the classical machine learning models. The CNN-LSTM model trained on data for thirteen key parameters achieved remarkable performance metrics, with an R^2 of 0.94 and an MSE of 0.12. The proposed model showed reliability in predicting DO concentration, particularly for water quality data from in-situ sensors. The results indicated that a hybrid CNN-LSTM model can predict water quality parameters such as DO concentration with improved accuracy compared to the classical models. The results of feature importance analysis using XGBoost showed that pH, with a relative importance of 0.76, was the most significant predictor. Temperature was the second significant predictor, with a relative importance of 0.12. These results suggest that real-time prediction of DO concentration with acceptable accuracy is possible even when data from its physical sensor is missing, as long as information from two other in-situ sensors is available. This research not only contributes to the advancement of predictive modeling in environmental monitoring but also underscores the potential of hybrid deep learning models to enhance the accuracy

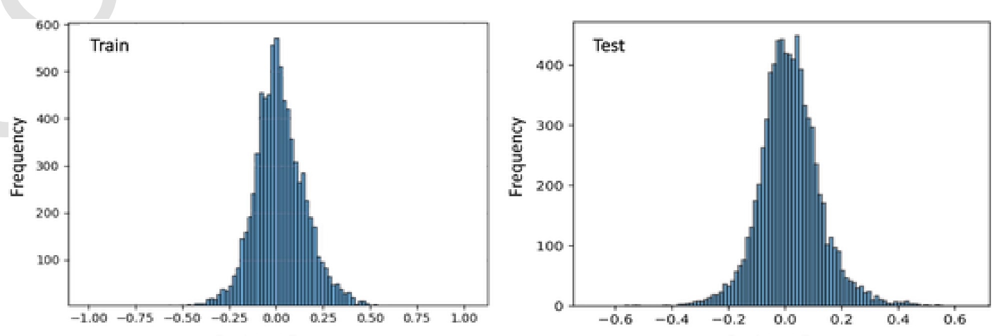

Fig. 8. Residual error vs. frequency for the predictions of the DO concentration by the hybrid CNN-LSTM model based on the training and test datasets.

and reliability of such predictions. This predictive method is an effective tool to support decision-making for water quality management and conservation efforts through autonomous monitoring of water quality. CNN-LSTM models are computationally intensive and require significant processing power, especially with real-time water quality data. Developing hybrid CNN-LSTM models suitable for microcontrollers is an important research area that supports autonomous monitoring of water quality.

CRediT authorship contribution statement

Majid Bagheri: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Karim Bagheri:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Nakisa Farshforoush:** Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Data curation. **Antonio Velazquez:** Supervision, Project administration, Funding acquisition, Formal analysis, Data curation. **Ying Liu:** Supervision, Software, Funding acquisition, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Department of Energy Minority Serving Institution Partnership Program (EM-MSIPP) managed by the Savannah River National Laboratory under BSRA contract TOA Number: 0000663608 and National Science Foundation under Award Number 2348805.

Data availability

The data for this study is available in Excel format as *Supplementary Material*.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2024.106595.

References

[1] M. Guo, R. Noori, S. Abolfathi, Microplastics in freshwater systems: dynamic behaviour and transport processes, Resour. Conserv. Recycl. 205 (2024) 107578, https://doi.org/10.1016/j.resconrec.2024.107578.
[2] H. Tian, L. Wang, X. Zhu, M. Zhang, L. Li, Z. Liu, S. Abolfathi, Biodegradation of microplastics derived from controlled release fertilizer coating: selective microbial colonization and metabolism in plastisphere, Sci. Total Environ. 920 (2024) 170978, https://doi.org/10.1016/j.scitotenv.2024.170978.
[3] S. Kirschke, T. Avellán, I. Bärlund, J.J. Bogardi, L. Carvalho, D. Chapman, C.W. Dickens, K. Irvine, S. Lee, T. Mehner, Capacity challenges in water quality monitoring: understanding the role of human development, Environ. Monit. Assess. 192 (2020) 1–16, https://doi.org/10.1007/s10661-020-8224-3.
[4] K. Murphy, B. Heery, T. Sullivan, D. Zhang, L. Paludetti, K.T. Lau, D. Diamond, E. Costa, F. Regan, A low-cost autonomous optical sensor for water quality monitoring, Talanta 132 (2015) 520–527, https://doi.org/10.1016/j.talanta.2014.09.045.
[5] O. Korostynska, A. Mason, A. Al-Shamma'a, Monitoring pollutants in wastewater: traditional lab based versus modern real-time approaches, in: S.C. Mukhopadhyay, A. Mason (Eds.), Smart Sensors for Real-time Water Quality Monitoring, Springer, Berlin, Heidelberg, 2013, pp. 1–24, https://doi.org/10.1007/978-3-642-37006-9_1.
[6] B.S. Pattnaik, A.S. Pattanayak, S.K. Udgata, A.K. Panda, Machine learning based soft sensor model for BOD estimation using intelligence at edge, Complex Intell. Syst. 7 (2021) 961–976, https://doi.org/10.1007/s40747-020-00259-9.
[7] J. Qiao, Z. Hu, W. Li, Soft measurement modeling based on chaos theory for biochemical oxygen demand (BOD), Water 8 (2016) 581, https://doi.org/10.3390/w8120581.
[8] F. Curreri, G. Fiumara, M.G. Xibilia, Input selection methods for soft sensor design: a survey, Future Internet 12 (2020) 97, https://doi.org/10.3390/fi12060097.
[9] A.K. Pani, V.K. Vadlamudi, H.K. Mohanta, Development and comparison of neural network based soft sensors for online estimation of cement clinker quality, ISA Trans. 52 (2013) 19–29, https://doi.org/10.1016/j.isatra.2012.07.004.
[10] M. Bagheri, N. Farshforoush, K. Bagheri, A.I. Shemirani, Applications of artificial intelligence technologies in water environments: from basic techniques to novel tiny machine learning systems, Process. Saf. Environ. Prot. 180 (2023) 10–22, https://doi.org/10.1016/j.psep.2023.09.072.
[11] D.H. Kwon, S.M. Hong, A. Abbas, J. Pyo, H.-K. Lee, S.-S. Baek, K.H. Cho, Inland harmful algal blooms (HABs) modeling using internet of things (IoT) system and deep learning, Environ. Eng. Res. 28 (2023), https://doi.org/10.4491/eer.2021.280.
[12] J. Kim, H. Kim, K. Kim, J.M. Ahn, Assessing the applicability of deep-learning method for predicting cyanobacteria in a regulated river, J. Environ. Eng. 150 (2024) 04024012, https://doi.org/10.1061/JOEEDU.EEENG-7427.
[13] G. Elkiran, V. Nourani, S. Abba, Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach, J. Hydrol. 577 (2019) 123962, https://doi.org/10.1016/j.jhydrol.2019.123962.
[14] T.M. Tung, Z.M. Yaseen, A survey on river water quality modelling using artificial intelligence models: 2000–2020, J. Hydrol. 585 (2020) 124670, https://doi.org/10.1016/j.jhydrol.2020.124670.
[15] M. Bagheri, A. Bazvand, M. Ehteshami, Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts, J. Clean. Prod. 149 (2017) 784–796 doi:2010.1016/j.jclepro.2017.02.157.
[16] A.Y. Sun, B.R. Scanlon, How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions, Environ. Res. Lett. 14 (2019) 073001, https://doi.org/10.1088/1748-9326/ab1b7d.
[17] H. Kamyab, T. Khademi, S. Chelliapan, M. SaberiKamarposhti, S. Rezania, M. Yusuf, M. Farajnezhad, M. Abbas, B.H. Jeon, Y. Ahn, The latest innovative avenues for the utilization of artificial intelligence and big data analytics in water resource management, Res. Eng. Des. 101566 (2023), https://doi.org/10.1016/j.rineng.2023.101566.
[18] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020) 115454, https://doi.org/10.1016/j.watres.2019.115454.
[19] Á.F. Gambín, E. Angelats, J.S. González, M. Miozzo, P. Dini, Sustainable marine ecosystems: deep learning for water quality assessment and forecasting, IEEE Access 9 (2021) 121344–121365, https://doi.org/10.1109/ACCESS.2021.3109216.
[20] A. Saeed, A. Alsini, D. Amin, Water quality multivariate forecasting using deep learning in a West Australian estuary, Environ. Model Softw. 171 (2024) 105884, https://doi.org/10.1016/j.envsoft.2023.105884.
[21] C. Shen, A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res. 54 (2018) 8558–8593, https://doi.org/10.1029/2018WR022643.
[22] A. Mosavi, S. Ardabili, A.R. Varkonyi-Koczy, List of deep learning models, in: International Conference on Global Research and Education, Springer, 2019, pp. 202–214, https://doi.org/10.1007/978-3-030-36841-8_20.
[23] D. Lee, B. Lee, S. Chae, S. Kwon, S. Kang, Integrated Prediction of Harmful Algal Bloom Using Cnn Deep Learning Technology, 4060909, 2020 https://doi.org/10.2139/ssrn.4060909, Available at SSRN.
[24] S.M. Ahmadi, S. Balahang, S. Abolfathi, Predicting the hydraulic response of critical transport infrastructures during extreme flood events, Eng. Appl. Artif. Intell. 133 (2024) 108573, https://doi.org/10.1016/j.engappai.2024.108573.
[25] P. Liu, J. Wang, A.K. Sangaiah, Y. Xie, X. Yin, Analysis and prediction of water quality using LSTM deep neural networks in IoT environment, Sustain 11 (2019) 2058, https://doi.org/10.3390/su11072058.
[26] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, X. Ma, B.L. Marrone, Z.J. Ren, J. Schrier, Machine learning: new ideas and tools in environmental science and engineering, Environ. Sci. Technol. 55 (2021) 12741–12754, https://doi.org/10.1021/acs.est.1c01339.
[27] J. Jia, W. Liang, Y. Liang, A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing, arXiv, 2023 https://doi.org/10.48550/arXiv.2312.05589, preprint arXiv:2312.05589.
[28] S.-S. Baek, J. Pyo, J.A. Chun, Prediction of water level and water quality using a CNN-LSTM combined deep learning approach, Water 12 (2020) 3399, https://doi.org/10.3390/w12123399.
[29] J.-J. Zhu, S. Borzooei, J. Sun, Z.J. Ren, Deep learning optimization for soft sensing of hard-to-measure wastewater key variables, ACS ES&T Eng. 2 (2022) 1341–1355, https://doi.org/10.1021/acsestengg.1c00469.
[30] S. Heo, K. Nam, J. Loy-Benitez, C. Yoo, Data-driven hybrid model for forecasting wastewater influent loads based on multimodal and ensemble deep learning, IEEE Trans. Ind. Inform. 17 (2020) 6925–6934, https://doi.org/10.1109/TII.2020.3039272.
[31] Y. Chen, W. Chen, S. Chandra Pal, A. Saha, I. Chowdhuri, B. Adeli, S. Janizadeh, A.A. Dineva, X. Wang, A. Mosavi, Evaluation efficiency of hybrid deep learning

algorithms with neural network decision tree and boosting methods for predicting groundwater potential, Geocarto Int. 37 (2022) 5564–5584, https://doi.org/10.1080/10106049.2021.1920635.
[32] M.D. Petkewich, J.M. Clark, L.J. Bodkin, B.J. Huffman, Water-Quality Measurements in Savannah River, Savannah, Georgia, 2020: U.S. Geological Survey Data Release, 2021, https://doi.org/10.5066/P94YZQQF.
[33] I. YSI, 6-Series Multiparameter Water Quality Sondes User Manual, 2012.
[34] W. Zhu, Z.T. Webb, K. Mao, J. Romagnoli, A deep learning approach for process data visualization using t-distributed stochastic neighbor embedding, Ind. Eng. Chem. Res. 58 (2019) 9564–9575, https://doi.org/10.1021/acs.iecr.9b00975.
[35] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html.
[36] J. Kim, C. Lee, J. Park, N. Kim, S.-L. Kim, J. Baek, Y.-S. Chung, K. Kim, Comparison of various drought resistance traits in soybean (Glycine max L.) based on image analysis for precision agriculture, Plant 12 (2023) 2331, https://doi.org/10.3390/plants12122331.
[37] A. Zien, N. Krämer, S. Sonnenburg, G. Rätsch, The feature importance ranking measure, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7–11, 2009, Proceedings, Part II 20, Springer, 2009, pp. 694–709, https://doi.org/10.1007/978-3-642-04174-7_45.
[38] C. Wade, K. Glynn, Hands-on Gradient Boosting With XGBoost and Scikit-Learn: Perform Accessible Machine Learning and Extreme Gradient Boosting With Python, Packt Publishing Ltd, UK, 2020.
[39] J.M. Ahn, J. Kim, K. Kim, Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting, Toxins 15 (2023) 608, https://doi.org/10.3390/toxins15100608.
[40] G. Hackeling, Mastering Machine Learning With Scikit-Learn, Packt Publishing Ltd, UK, 2017.
[41] R. Shyam, Convolutional neural network and its architectures, J. Comput. Technol. Appl. 12 (2021) 6–14.
[42] W. Hao, W. Yizhou, L. Yaqin, S. Zhili, The role of activation function in CNN, in: 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE, 2020, pp. 429–432, https://doi.org/10.1109/ITCA52113.2020.00096.
[43] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A.K. Dutta, S. Almotairi, A comparison of pooling methods for convolutional neural networks, Appl. Sci. 12 (2022) 8643, https://doi.org/10.3390/app12178643.
[44] M. Sewak, M.R. Karim, P. Pujari, Practical Convolutional Neural Networks: Implement Advanced Deep Learning Models Using Python, Packt Publishing Ltd, UK, 2018.
[45] A. Graves, Long short-term memory, in: Supervised Sequence Labelling With Recurrent Neural Networks, Springer, Berlin, Heidelberg, 2012, pp. 37–45, https://doi.org/10.1007/978-3-642-24797-2_4.
[46]
[47] M. Adnan, A.A.S. Alarood, M.I. Uddin, I. ur Rehman, Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models, PeerJ Comput. Sci. 8 (2022) e803, https://doi.org/10.7717/peerj-cs.803.
[48] S. Borzooei, L. Scabini, G. Miranda, S. Daneshgar, L. Deblieck, O. Bruno, P. De Langhe, B. De Baets, I. Nopens, E. Torfs, Evaluation of activated sludge settling characteristics from microscopy images with deep convolutional neural networks and transfer learning, J. Water Process Eng. 64 (2024) 105692, https://doi.org/10.1016/j.jwpe.2024.105692.
[49] M. Habib, J. O'Sullivan, S. Abolfathi, M. Salauddin, Enhanced wave overtopping simulation at vertical breakwaters using machine learning algorithms, PLoS One 18 (2023) e0289318, https://doi.org/10.1371/journal.pone.0289318.
[50] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, Peerj Comput. Sci. 7 (2021) e623. http://dx.doi.org/https://doi.org/10.7717/peerj-cs.623.
[51] S.A. Mirbagheri, M. Bagheri, S. Boudaghpour, M. Ehteshami, Z. Bagheri, Performance evaluation and modeling of a submerged membrane bioreactor treating combined municipal and industrial wastewater using radial basis function artificial neural networks, J. Environ. Health Sci. Eng. 13 (2015) 1–15, https://doi.org/10.1186/s40201-015-0172-4.
[52] H. Liu, J. Yang, M. Ye, S.C. James, Z. Tang, J. Dong, T. Xing, Using t-distributed stochastic neighbor embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data, J. Hydrol. 597 (2021) 126146, https://doi.org/10.1016/j.jhydrol.2021.126146.
[53] R. Harvey, L. Lye, A. Khan, R. Paterson, The influence of air temperature on water temperature and the concentration of dissolved oxygen in Newfoundland Rivers, Can. Water Resour. J. 36 (2011) 171–192, https://doi.org/10.4296/cwrj3602849.
[54] J. Abowei, Salinity, dissolved oxygen, pH and surface water temperature conditions in Nkoro River, Niger Delta, Nigeria, Adv. J. Food Sci. Technol. 2 (2010) 36–40.
[55] J.J. Dabrowski, A. Rahman, A. George, Prediction of dissolved oxygen from pH and water temperature in aquaculture prawn ponds, in: Proceedings of the Australasian Joint Conference on Artificial Intelligence-Workshops, 2018, pp. 2–6, https://doi.org/10.1145/3314487.3314488.
[56] S.-Q. Dotse, I. Larbi, A.M. Limantol, L.C. De Silva, A review of the application of hybrid machine learning models to improve rainfall prediction, Model. Earth Sys. Environ. 10 (2024) 19–44, https://doi.org/10.1007/s40808-023-01835-x.
[57] A. Wehenkel, J. Behrmann, H. Hsu, G. Sapiro, G. Louppe, J.-H. Jacobsen, Robust Hybrid Learning With Expert Augmentation, arXiv, 2022 https://doi.org/10.48550/arXiv.2202.03881, preprint arXiv:2202.03881.
[58] M. Bagheri, S.A. Mirbagheri, M. Ehteshami, Z. Bagheri, A.M. Kamarkhani, Analysis of variables affecting mixed liquor volatile suspended solids and prediction of effluent quality parameters in a real wastewater treatment plant, Desalin. Water Treat. 57 (2016) 21377–21390, https://doi.org/10.1080/19443994.2015.1125796.
[59] J. Baum, H. Kanagawa, A. Gretton, A kernel stein test of goodness of fit for sequential models, in: International Conference on Machine Learning, PMLR, 2023, pp. 1936–1953.
[60] S. Heddam, Fuzzy neural network (EFuNN) for modelling dissolved oxygen concentration (DO), in: C. Kahraman, İ.U. Sari (Eds.), Intelligence Systems in Environmental Management: Theory and Applications, Springer Cham, 2017, pp. 231–253, https://doi.org/10.1007/978-3-319-42993-9_11.
[61] Y. Liu, Q. Zhang, L. Song, Y. Chen, Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction, Comput. Electron. Agric. 165 (2019) 104964, https://doi.org/10.1016/j.compag.2019.104964.
[62] F.H. Garabaghi, S. Benzer, R. Benzer, Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach, Environ. Monit. Assess. 195 (2023) 879, https://doi.org/10.1007/s10661-023-11492-3.
[63] R. Dehghani, H. Torabi Poudeh, Z. Izadi, Dissolved oxygen concentration predictions for running waters with using hybrid machine learning techniques, Model. Earth Sys. Environ. (2021) 1–15, https://doi.org/10.1007/s40808-021-01253-x.
[64] M.H. Ahmed, L.-S. Lin, Dissolved oxygen concentration predictions for running waters with different land use land cover using a quantile regression forest machine learning technique, J. Hydrol. 597 (2021) 126213, https://doi.org/10.1016/j.jhydrol.2021.126213.
[65] W. Zhi, D. Feng, W.-P. Tsai, G. Sterle, A. Harpold, C. Shen, L. Li, From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? Environ. Sci. Technol. 55 (2021) 2357–2368, https://doi.org/10.1021/acs.est.0c06783.
[66] S.V. Moghadam, A. Sharafati, H. Feizi, S.M.S. Marjaie, S.B.H.S. Asadollah, D. Motta, An efficient strategy for predicting river dissolved oxygen concentration: application of deep recurrent neural network model, Environ. Monit. Assess. 193 (2021) 1–18, https://doi.org/10.1007/s10661-021-09586-x.
[67] D. Pan, Y. Zhang, Y. Deng, J. Van Griensven Thé, S.X. Yang, B. Gharabaghi, Dissolved oxygen forecasting for Lake Erie's Central Basin using hybrid long short-term memory and gated recurrent unit networks, Water 16 (2024) 707, https://doi.org/10.3390/w16050707.
[68] Z. Zheng, H. Ding, Z. Weng, L. Wang, Research on a multiparameter water quality prediction method based on a hybrid model, Ecol. Inform. 76 (2023) 102125, https://doi.org/10.1016/j.ecoinf.2023.102125.
[69] M. Bagheri, S.A. Mirbagheri, Z. Bagheri, A.M. Kamarkhani, Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach, Process. Saf. Environ. Prot. 95 (2015) 12–25, https://doi.org/10.1016/j.psep.2015.02.008.