Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 6644681, 10 pages
https://doi.org/10.1155/2023/6644681

Research Article

Water Quality Prediction Based on Hybrid Deep Learning Algorithm

Bhagavathi Perumal,[1] Niveditha Rajarethinam,[2] Anusuya Devi Velusamy,[3] and Venkatesa Prabhu Sundramurthy[4]

[1]*Department of Civil Engineering, Sri Sairam Engineering College, Chennai 600044, Tamil Nadu, India*
[2]*Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, Tamil Nadu, India*
[3]*Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil 626126, Tamil Nadu, India*
[4]*Center of Excellence for Bioprocess and Biotechnology, Department of Chemical Engineering, College of Biological and Chemical Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia*

Correspondence should be addressed to Venkatesa Prabhu Sundramurthy; venkatesa.prabhu@aastu.edu.et

Received 19 April 2023; Revised 10 September 2023; Accepted 3 October 2023; Published 7 November 2023

Academic Editor: Manoj Khandelwal

Pollution from many different sources severely affects the quality of our water supply. Over the past few years, a large number of online water quality monitoring stations have been used to gather time series data on water quality monitoring. These numbers are the foundation for deep learning techniques for forecasting water quality. In particular, typical deep learning approaches struggle to accurately estimate water quality in the presence of net promoter system (NPS) contamination. To overcome this shortcoming, a new deep learning model called long short-term memory (LSTM)–gray wolf optimization (GWO)–fish swarm optimization (FSO) was developed to enhance the precision of water quality prediction with NPS pollution. The well-established model may remedy the mechanism models' inability to foretell changes in water quality on a minute-by-minute basis. Thamirabarani river watershed was used for the model's application. Based on experimental data, the suggested model outperformed the mechanism model and the LSTM model in predicting extreme values. Maximum relative errors in anticipated against observed dissolved oxygen, chemical oxygen demand, and NH_3—N values were 7.58%, 18.45%, and 22.25%, respectively. In comparison to the artificial neural network (ANN), back propagation neural network (BPNN), and recurrent neural network (RNN) models, the created LSTM–GWO–FSO model was shown to have greater computational performance (RNN). LSTM–GWO–FSO outperformed ANN, BPNN, and RNN regarding R^2 of 3.1%–38.4% improvements. The suggested approach may provide a fresh perspective when predicting water quality in the presence of NPS contamination.

1. Introduction

Water pollution is the contamination of water bodies, such as lakes, rivers, oceans, and groundwater, by various substances, including chemicals, pathogens, and physical debris. Water pollution can occur naturally or as a result of human activities. Human existence depends on surface water, yet it is a finite and nonrenewable resource. However, human industry and urbanization growth are wreaking havoc on the natural environment. Threats to human health posed by the persistent contamination and degradation of the surface water environment are significant [1, 2]. Additionally, the issues above are made much worse by the degradation of the surface water ecosystem [3]. Thus, monitoring and making projections about surface water quality is crucial. Based on the water quality prediction exercise results, we may deduce that we must look to past environmental indicators to warn us of impending ecological contamination [4]. It is challenging to reliably anticipate future water quality indicator values owing to the nonlinearity and nonstationarity of water quality data. As a promising new avenue for improving water quality forecasting, artificial intelligence (AI) technology is worth exploring. The use of AI to improve human lives is spreading to more and more disciplines [5, 6]. The main areas of interest for water quality prediction using the gray system theory, neural networks, statistical analysis techniques, and time series models are lakes, rivers, reservoirs, estuaries, and other vast expanses of water.

Machine learning has been used for various water treatment and management issues, such as real-time monitoring, forecasting, locating the source of pollutants, estimating their concentrations, allocating water resources, and improving water treatment technologies. The wastewater from cities and industries is the primary cause of water pollution in urban areas [7]. Machine-learning applications in surface water quality studies have recently emerged as a focal point [8, 9]. Several approaches for analyzing and forecasting surface water quality have been developed. Many attempts have been made to enhance the accuracy of predictions made by machine-learning models. Gathering relevant data is a cornerstone of creating effective machine-learning models [10–12]. It is possible to utilize data from integrated and periodic water quality monitoring as reference points for managing the water system. Environmental authorities commonly use conventional ecological monitoring techniques. Traditional methods of in situ monitoring have practical limitations [13]. The demands of continuous, widespread water quality monitoring are within the capabilities of remote sensing technology. In addition, they may provide light on the elusive migratory and dispersion patterns of contaminants, which are detectable only by these means.

As Sagan et al. [14] discovered that experiment-based machine learning paves the way for advanced optimization using real-time sensor and satellite data monitoring. Standard models were outperformed by those based on partial least squares regression, support vector regression (SVR), and deep neural networks. Some water quality indicators, such as pathogen content, are not optically active and do not have high-spatial-resolution hyperspectral data. Still, they may be approximated indirectly using other measurable data. Wu et al. [15] created an attentional neural network built on top of a convolutional neural network (CNN) to distinguish between uncontaminated and contaminated water. They validated the efficiency of their attentional neural network by conducting many comparison tests using a collection of images of the water's surface. CNN's use of the reflectance picture as the only input eliminates the need for feature engineering and fine-tuning, which is a significant benefit.

In recent years, multivariate statistical approaches have found widespread use in groundwater quality analysis. Standard methods include principal component and cluster analysis [16]. Support vector machines, Decision tree, random forest (RF), and artificial neural network (ANN) are just some machine-learning techniques used to evaluate groundwater quality. Research in this area has focused chiefly on evaluating the efficacy of various machine-learning algorithms for assessing groundwater quality to select the most appropriate algorithms for a given set of circumstances. It is common to practice building mechanism models for water quality prediction by first gaining knowledge of the relevant physical processes and components. Parameters of mechanism models have rigorous physical interpretation [17], implying its usefulness. Challenges in parameter calibration, complex modeling frameworks, uncertain model parameters, and high computing cost limit their use in watershed water quality prediction. In addition to being cumbersome to install and calibrate, mechanism models have a poor reputation. Using mechanistic models, the reduction in water quality caused by net promoter system (NPS) pollution is difficult to anticipate in real-time or over short periods [18].

To evaluate the fundamental factors impacting semiarid groundwater and how they affect areas of high-quality groundwater in Tabriz City, Iran, Jeihouni et al. [19] examined the performance of five data mining algorithms: RF, chi-square automatic interaction detector, iterative dichotomized, and regular decision tree. Self-organizing neural networks and fuzzy c-means clustering are combined. Lee et al. [20] assessed the geographical pattern of urban groundwater quality in Seoul, South Korea. Using a self-organizing map technique, they separated the groundwater samples into three categories depending on their contamination level. Then, they used this information to examine the pollution-driven process in terms of its geographical manifestation. Geographic information system methods have been widely employed to improve groundwater pollution detection to create quality maps of underground water [14, 21–23].

This work aims to predict water quality metrics with high precision, introducing a hybrid model using deep learning and optimization methods. The Thamirabarani river basin was used as the study's focal point. The high-precision prediction methods of several water quality indices (WQIs) were examined and proposed; these included pH, dissolved oxygen (DO), chemical oxygen demand (COD), NH_3—N, and water quality guidelines. Finally, an long short-term memory (LSTM)–gray wolf optimization (GWO)–fish swarm optimization (FSO) model was developed to address the problem of WQI prediction over the long term.

2. Materials and Methods

An important river in southern India, the Thamirabarani flows through the state of Tamil Nadu. Several potential influences on the Thamirabarani river's water quality must be evaluated to provide an accurate prediction of the river's water quality. Researchers and officials may take water samples from the Thamirabarani river and test them for nutrient, bacterial, metal, and other pollution levels to make educated predictions about the river's water quality. Changes in land use, weather patterns, and other variables may all be modeled and simulated to determine their potential future effects on water quality. Thamirabarani river water quality may be protected and enhanced with the help of this data, which will be used to guide policy decisions. The study area of the Thamirabarani river basin is shown in Figure 1.

2.1. Method. The water quality evaluation down to the molecular level uses a WQI, which considers a wide range of physical, chemical, and biological factors. Typically, the index will offer a single score to sum up the water quality, making it more straightforward for policymakers, researchers, and the general public to comprehend and discuss the status of the water. Location, water source, and end-use all

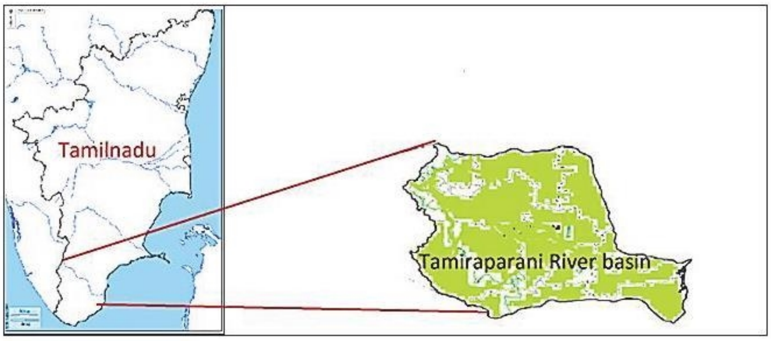


FIGURE 1: Thamirabarani river basin study area.

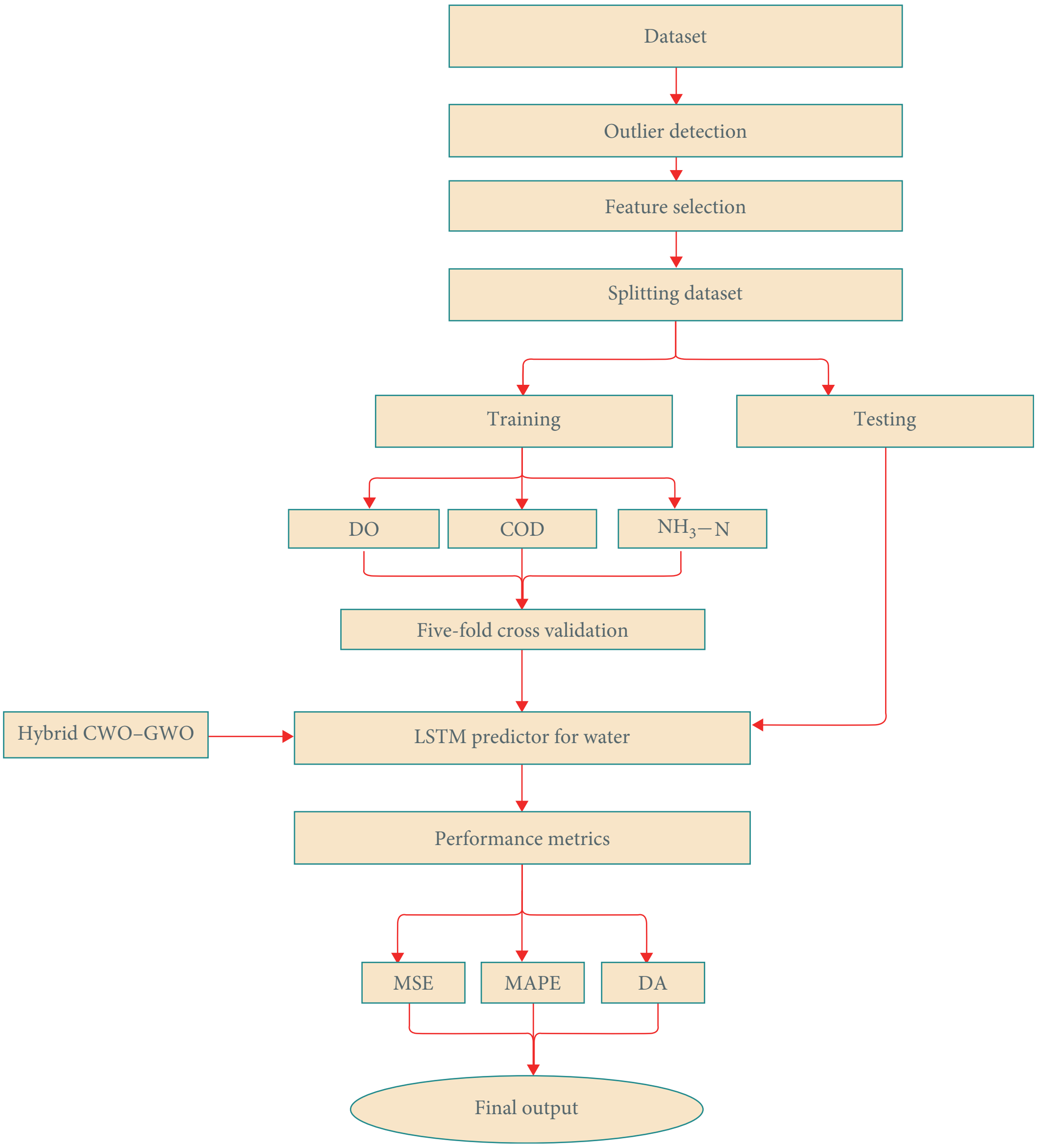


FIGURE 2: Methodology adopted in this study.

have a role in determining WQI values. While some indices are developed with drinking water in mind, others are used on lakes, rivers, and other open water bodies. WQIs may be calculated using a wide range of factors and scoring methods, all depending on the individual situation and the end users relying on the index. The methodology adopted in the study is shown in Figure 2. The testing and training data of the study are presented in Figure 3.

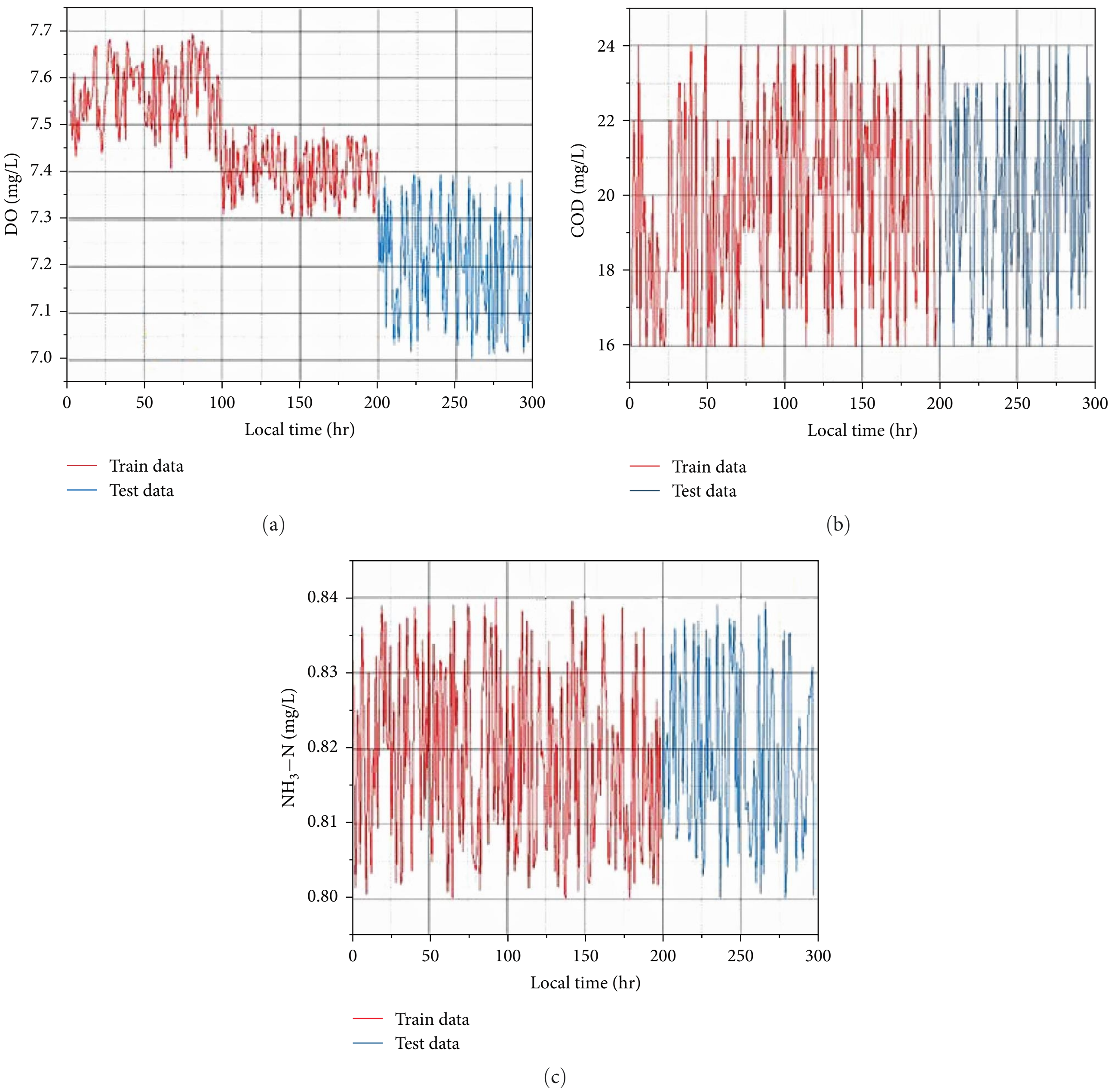


Figure 3: Testing and training dataset of the study.

2.2. GWO. The social structure and hunting techniques of gray wolves in the wild inspired the GWO algorithm, which optimizes throughout a population. The GWO method is a metaheuristic optimization approach with several potential applications outside traditional optimization settings. The GWO algorithm begins by seeding the world with a group of search agents that stand in for gray wolves in the wild. A search agent, also known as a wolf, is given a location in the search space corresponding to a particular feasible answer to the optimization issue. A social hierarchy similar to that of a pack of wolves, alpha, beta, delta, and omega members governs the search agents' interactions. The GWO method is fast, efficient, and successful in solving various optimization problems. It is easy to implement and does not require familiarity with the optimization issue.

2.3. FSO. FSO is a population-based optimization technique based on observations of schooling fish. Individual learning, social learning, and global learning all play a role in determining where each fish goes in the search space. The social learning component depicts the fish's propensity to observe and mimic the actions of its peers. In contrast, the individual learning factor shows the fish's tendency to explore the search area independently. The global learning factor reflects the fish's propensity to stick with the best answer it's discovered so far in the search space. The FSO method quickly and efficiently finds optimum or near-optimal solutions to various optimization problems. It is easy to implement and does not require familiarity with the optimization issue.

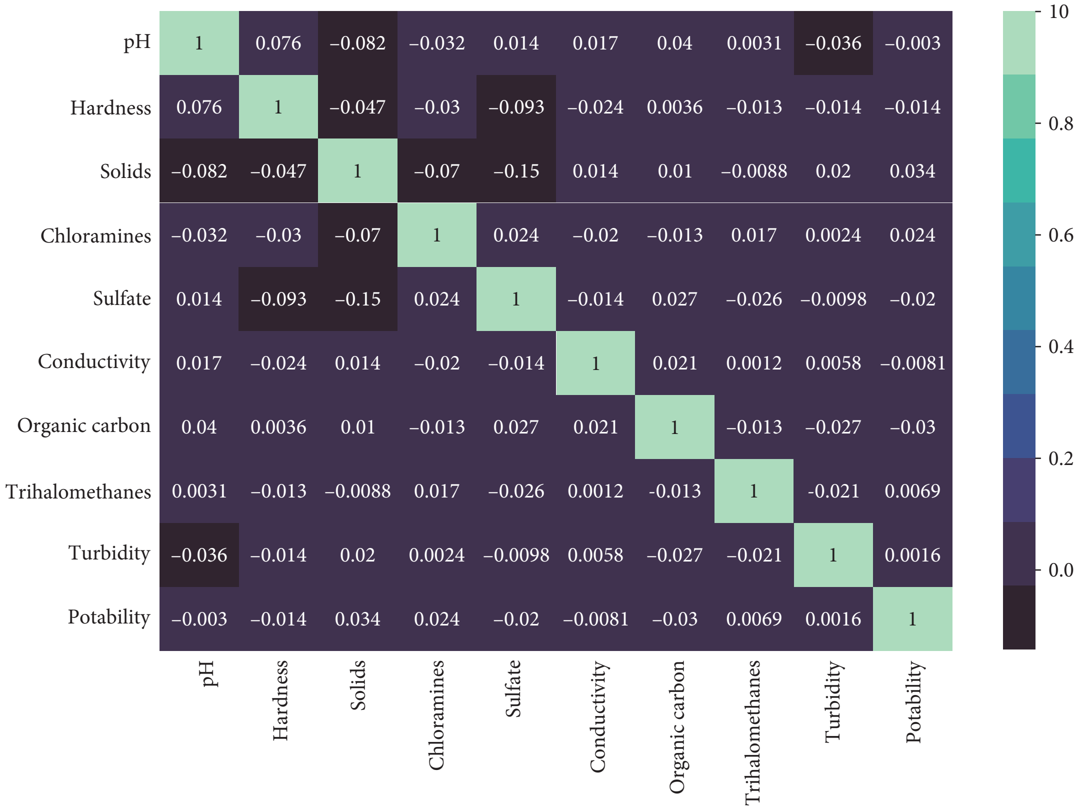


Figure 4: Correlations plot of the feature selections.

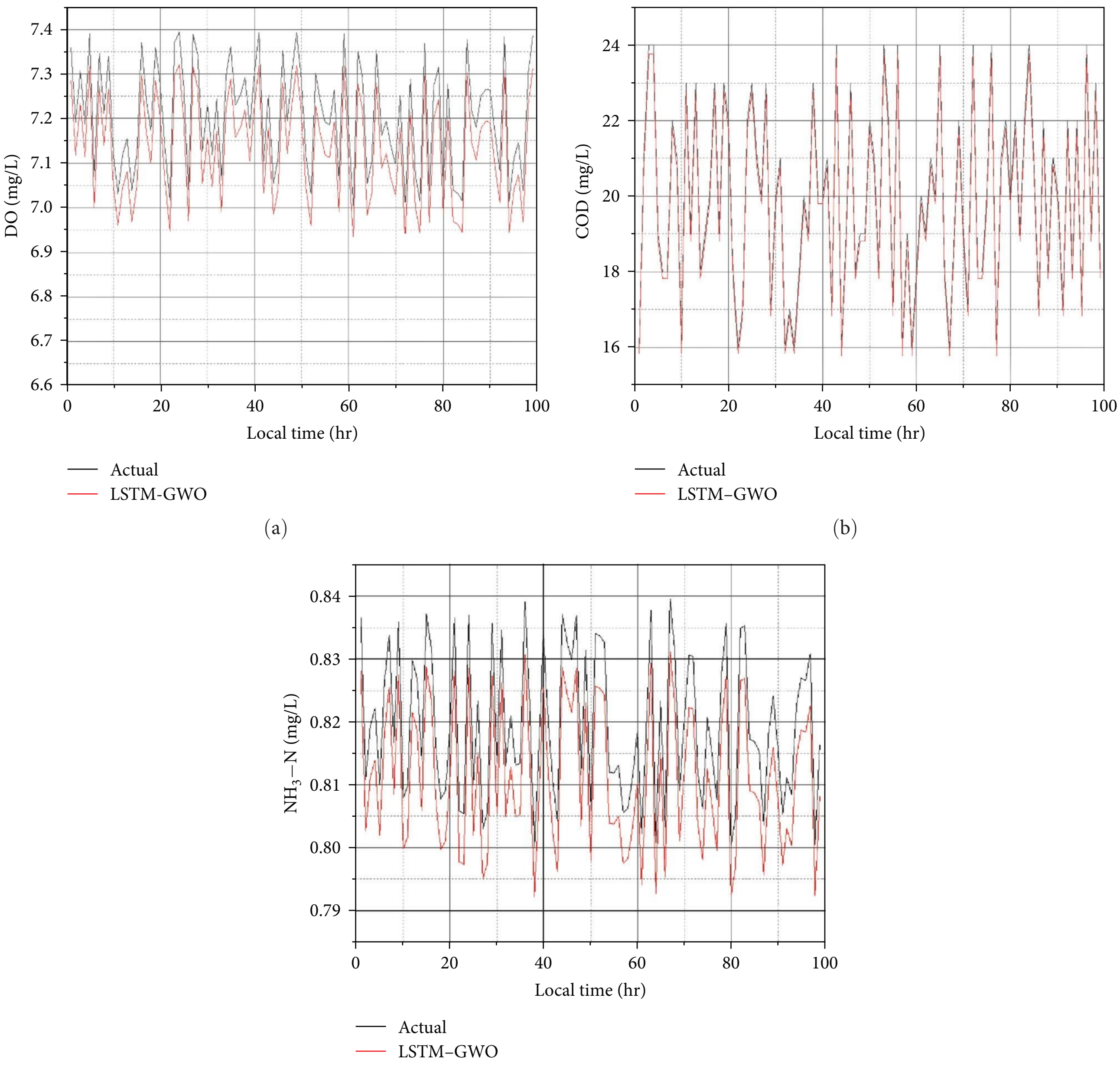
3. Result and Discussion

The term “heat map” refers to a specific kind of graphical representation used in data analysis and presentation that uses color to denote numerical values. Machine learning and AI programs often use heat maps to help visualize the distribution of data values and spot patterns and trends. Figure 4 shows the correlation plot of the features in the Thamirabarani river basin area. The study analyzed the water quality predictions.

In this section, the constructed LSTM–GWO–FSO model’s performance is evaluated using water quality indicators such as DO, COD, and ammonia nitrogen (NH_3—N). Hourly DO, COD, and NH_3—N concentration values were projected from May 2021 to December 2022. we were using online monitoring data collected between May 2021 and December 2022. Before being fed into the model, the input data underwent a series of preparation steps. During this phase, we eliminated outliers using the three principles, replaced missing data with average daily values, and removed incorrect information. As a bonus, the input dataset was normalized using the *Z*-score standardization approach to speed up training and boost the model’s prediction accuracy. Figure 5 shows actual and LSTM–GWO predicted values of (a) DO, (b) COD, and (c) NH3—N.

The best possible LSTM–GWO–FSO model was achieved after training and convergence. The test dataset was utilized for model assessment to ensure the best model prediction performance. Figure 6 compares the model’s predicted values and the corresponding observed values between June 15, 2022, and June 28, 2022. DO, COD, and NH_3—N had the highest relative errors of 7.58%, 18.45%, and 22.25%, respectively. Some businesses’ hidden or leakage emissions may not have been accounted for in the original model, leading to unexpected shifts in water quality and a substantial TP forecast error. Figure 7 shows actual and LSTM–FSO predicted values of (a) DO, (b) COD, and (c) NH3—N.

When the concentration of NPS pollutants suddenly shifts due to harsh weather, standard deep learning algorithms are constrained by past data. They cannot adequately anticipate extreme values, whereas physical models can guarantee that the predictions are within a manageable range. The suggested model, LSTM–GWO–FSO, was evaluated on data not included in the training set to gauge its ability to predict outlying values. We compared the projected, observed, and calculated values for severe DO, COD, and NH_3—N concentrations during 30 days, and the findings are shown in Figure 6. The well-established LSTM–GWO–FSO model has attained high prediction accuracy after the linking mechanism model to forecast the extreme values daily. Previous research has shown that geographical features might affect the transit and dispersion process of NPS contamination in the watershed. However, conventional deep-learning approaches have largely ignored the effect of spatial factors on the transport-diffusion process of NPS contamination. The results of 60-hr predictions for DO, COD, NH_3—N, and TP from the LSTM–GWO–FSO model with geographical information and the LSTM model. We found that the LSTM–GWO–FSO model outperformed the LSTM model.



(a)

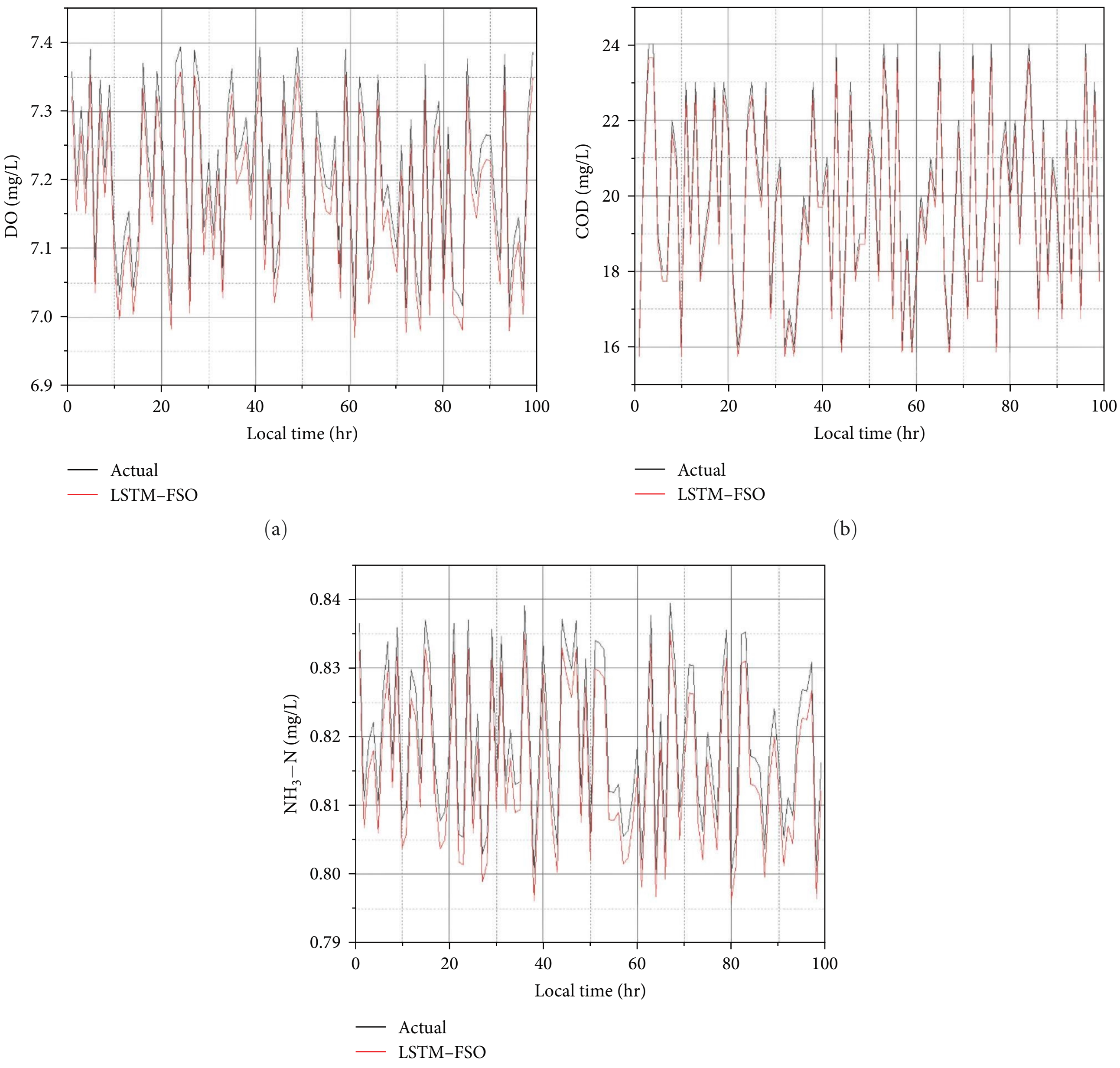
(b)

(c)

Figure 5: Actual and LSTM–GWO predicted values of (a) DO, (b) COD, and (c) NH_3—N.

By incorporating new state variables into time series data. In a recurrent neural network (RNN), the input to the hidden layer is composed of the previous output of the hidden layer as well as the current output of the network layer. Each of the hidden layer's nodes connects to every other node. This ensures that the preceding layers' results may impact the current concealed layer's output. This means that time series data is no problem for the RNN model. This investigation found that an RNN with three hidden layers and 30 neurons per layer was the most effective configuration. Table 1 displays the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) for four different models. A statistical analysis of the various models shows that the suggested sediment oxygen demand (SOD)–LSTM–GWO–FSO model provides the most accurate predictions. Out of four cutting-edge prediction models, ANN performed the poorest statistically. While the back propagation neural network (BPNN) model improved upon the ANN model regarding water quality prediction, it still lacked reliable forecasts. The model's performance might be enhanced by using neural networks like RNN and LSTM–GWO–FSO, which are optimized for processing sequence data. Table 2 shows that the well-tested models are trustworthy and consistent.

Figure 6 displays the ANN, BPNN, and RNN models and LSTM–GWO–FSO predictions for the DO, COD, and NH_3—N. While both the ANN and BPNN models could capture the overall upward trend in pollution levels over



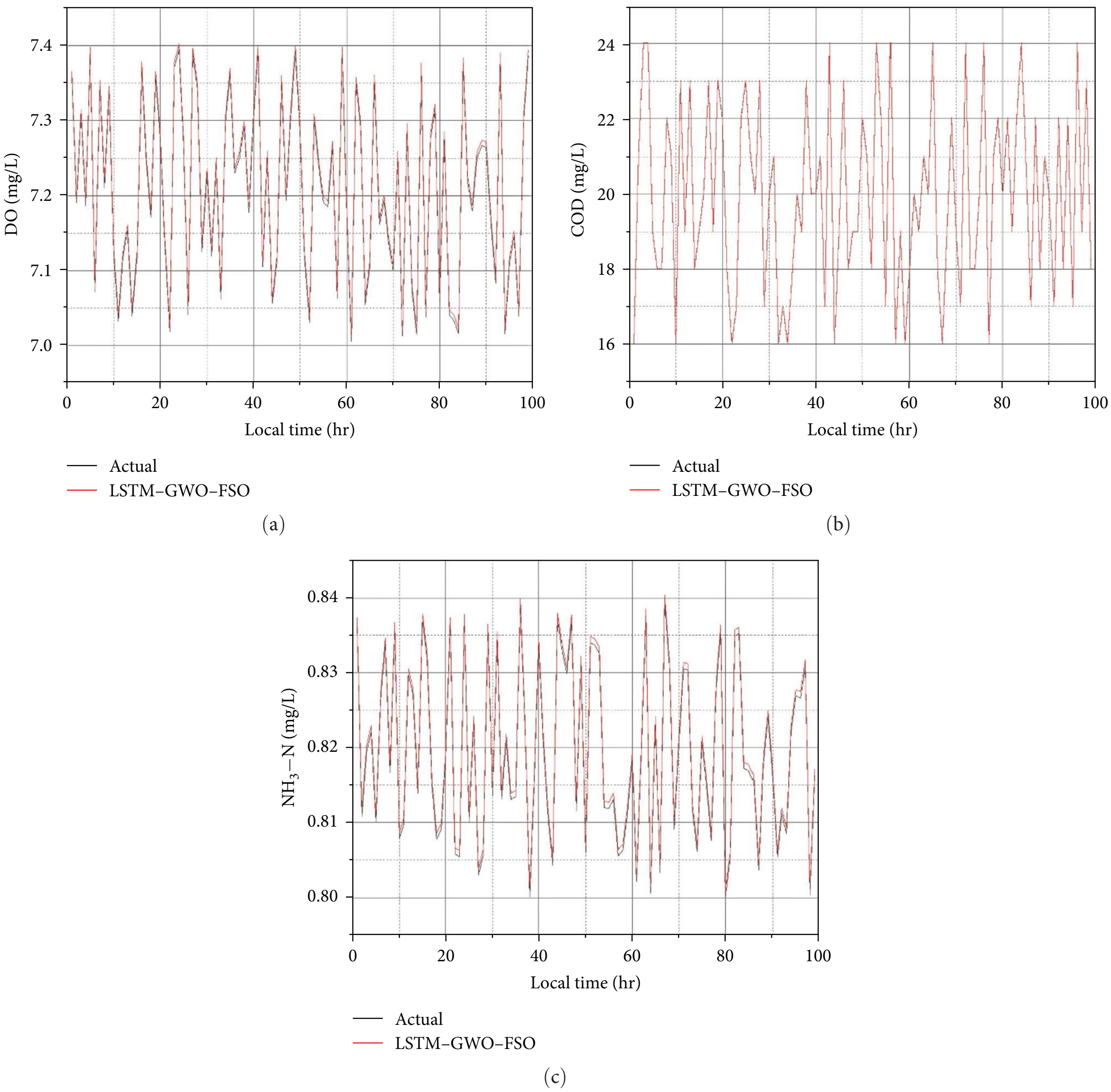
(a)

(b)

(c)

Figure 6: Actual and LSTM–GWO–FSO predicted values of (a) DO, (b) COD, and (c) NH_3—N.

time, they each had their own unique set of limitations when predicting outliers. Possible explanation: the ANN model uses autoregression and sliding averages. The ANN model's projections were not too far from the norm. When there was not a great deal of variation in the real value, ANN may be a better fit. A possible drop in extreme value prediction precision resulted from SVR's failure to account for the influence of time series data on prediction outcomes (the effect of one time period on the next). However, RNN and LSTM–GWO–FSO models performed well in predicting water quality. The water quality prediction accuracy of the LSTM–GWO–FSO model, which integrated a mechanical model with spatial data, was superior to that of the RNN model. Based on the assessment, LSTM–GWO–FSO outperformed ANN, BPNN, and RNN with an improved R^2 of 3.1%–38.4%.



(a)

(b)

(c)

Figure 7: Actual and LSTM–FSO predicted values of (a) DO, (b) COD, and (c) NH_3—N.

Table 1: Performance metrics of the variable methods.

Features	Method	RMSE	MAE	MAPE	R^2
Dissolved oxygen	ANN	0.247	0.199	0.175	0.91
	BPNN	0.179	0.139	0.119	0.90
	RNN	0.135	0.092	0.079	0.89
	LSTM–GWO–FSO	0.083	0.055	0.044	0.94
COD	ANN	0.059	0.038	0.291	0.902
	BPNN	0.049	0.029	0.191	0.91
	RNN	0.039	0.021	0.142	0.93
	LSTM–GWO–FSO	0.016	0.011	0.080	0.95
NH_3—N	ANN	0.015	0.014	0.598	0.92
	BPNN	0.013	0.011	0.219	0.915
	RNN	0.008	0.007	0.149	0.925
	LSTM–GWO–FSO	0.0055	0.0045	0.128	0.94

Table 2: Comparison of the existing algorithm with the present study.

Method	Performance metrics	MSE	MAPE	References
Ant bee colony BPNN	DO	0.683	9.43	[24]
	COD	9.43	27.07	
LSTM	DO	0.08	9.7	[6]
	COD	0.03	1.7	
LSTM–RF	DO	0.09	10.5	[5]
	COD	0.045	2	
LSTM–SOD–VGG	DO	0.225	0.014	[25]
	COD	0.051	0.046	
LSTM–GWO–FSO	DO	0.055	0.04	Present study
	COD	0.011	0.08	
	NH_3—N	0.0045	12.8	

4. Conclusions

A deep learning model composed of LSTM–GWO–FSO modules was developed for the research. The training set consisted of hydrometeorological data, pollutant parameters, an error sequence, and geographical features. LSTM's inaccuracy and water quality prediction resulted from incorrect pollutant concentrations. In addition to overcoming the difficulty of predicting extreme outcomes, the developed model also considered the impact of spatial components on water quality at different times and places. To measure how well the established model performed, it was compared to three state-of-the-art prediction models: ANN, BPNN, and RNN. The constructed LSTM–GWO–FSO model outperformed the ANN, BPNN, and RNN models in terms of computational performance (RNN). With an improved R^2 of 3.1%–38.4%, LSTM–GWO–FSO beat ANN, BPNN, and RNN. Improvements. In the context of estimating water quality in the face of NPS pollution, the proposed method may provide a new point of view.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors' Contributions

Bhagavathi Perumal has contributed to the conceptualization, Niveditha Rajarethinam has contributed to the supervision and methodology, Anusuya Devi Velusamy has done the investigations, and Venkatesa Prabhu Sundramurthy has contributed to the writing.

References

[1] C. J. Vörösmarty, P. B. McIntyre, M. O. Gessner et al., "Global threats to human water security and river biodiversity," *Nature*, vol. 467, pp. 555–561, 2010.

[2] J. Wang and A. V. Nguyen, "A review on data and predictions of water dielectric spectra for calculations of van der Waals surface forces," *Advances in Colloid and Interface Science*, vol. 250, pp. 54–63, 2017.

[3] Y. Sun, Z. Chen, G. Wu et al., "Characteristics of water quality of municipal wastewater treatment plants in China: implications for resources utilization and management," *Journal of Cleaner Production*, vol. 131, pp. 1–9, 2016.

[4] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, pp. 452–459, 2015.

[5] M. Rezaie-Balf, N. F. Attar, A. Mohammadzadeh et al., "Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: comparative assessment of a noise suppression hybridization approach," *Journal of Cleaner Production*, vol. 271, Article ID 122576, 2020.

[6] S. Zhao, S. Zhang, J. Liu et al., "Application of machine learning in intelligent fish aquaculture: a review," *Aquaculture*, vol. 540, Article ID 736724, 2021.

[7] R. Mohammadpour, S. Shaharuddin, C. K. Chang, N. A. Zakaria, A. A. Ghani, and N. W. Chan, "Prediction of water quality index in constructed wetlands using support vector machine," *Environmental Science and Pollution Research*, vol. 22, pp. 6208–6219, 2015.

[8] Tiyasha, T. M. Tung, and Z. M. Yaseen, "A survey on river water quality modelling using artificial intelligence models: 2000–2020," *Journal of Hydrology*, vol. 585, Article ID 124670, 2020.

[9] N. Sharma, R. Sharma, and N. Jindal, "Machine learning and deep learning applications—a vision," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 24–28, 2021.

[10] K. Elbaz, A. Zhou, and S.-L. Shen, "Deep reinforcement learning approach to optimize the driving performance of shield tunnelling machines," *Tunnelling and Underground Space Technology*, vol. 136, Article ID 105104, 2023.

[11] K. Elbaz, W. M. Shaban, A. Zhou, and S.-L. Shen, "Real time image-based air quality forecasts using a 3D-CNN approach with an attention mechanism," *Chemosphere*, vol. 333, Article ID 138867, 2023.

[12] J. Du, J. Zhang, D. Castro-Lacouture, and Y. Hu, "Lean manufacturing applications in prefabricated construction projects," *Automation in Construction*, vol. 150, Article ID 104790, 2023.

[13] W. Li, H. Fang, G. Qin et al., "Concentration estimation of dissolved oxygen in Pearl River basin using input variable selection and machine learning techniques," *Science of The Total Environment*, vol. 731, Article ID 139099, 2020.

[14] V. Sagan, K. T. Peterson, M. Maimaitijiang et al., “Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing,” *Earth-Science Reviews*, vol. 205, Article ID 103187, 2020.

[15] Y. Wu, X. Zhang, Y. Xiao, and J. Feng, “Attention neural network for water image classification under IoT environment,” *Applied Sciences*, vol. 10, no. 3, Article ID 909, 2020.

[16] Z. L. Hildenbrand, D. D. Carlton Jr., B. E. Fontenot et al., “A comprehensive analysis of groundwater quality in the Barnett Shale region,” *Environmental Science & Technology*, vol. 49, no. 13, pp. 8254–8262, 2015.

[17] Q. Zuo, Q. Wu, L. Yu, Y. Li, and Y. Fan, “Optimization of uncertain agricultural management considering the framework of water, energy and food,” *Agricultural Water Management*, vol. 253, Article ID 106907, 2021.

[18] J. Senent-Aparicio, P. Jimeno-Sáez, A. Bueno-Crespo, J. Pérez-Sánchez, and D. Pulido-Velázquez, “Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction,” *Biosystems Engineering*, vol. 177, pp. 67–77, 2019.

[19] M. Jeihouni, A. Toomanian, and A. Mansourian, “Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS,” *Water Resources Management*, vol. 34, pp. 139–154, 2020.

[20] K.-J. Lee, S.-T. Yun, S. Yu, K.-H. Kim, J.-H. Lee, and S.-H. Lee, “The combined use of self-organizing map technique and fuzzy c-means clustering to evaluate urban groundwater quality in Seoul metropolitan city, South Korea,” *Journal of Hydrology*, vol. 569, pp. 685–697, 2019.

[21] M. Zhu, J. Wang, X. Yang et al., “A review of the application of machine learning in water quality evaluation,” *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, 2022.

[22] K. P. Wai, M. Y. Chia, C. H. Koo, Y. F. Huang, and W. C. Chong, “Applications of deep learning in water quality management: a state-of-the-art review,” *Journal of Hydrology*, vol. 613, Part A, Article ID 128332, 2022.

[23] T. Rajaee, S. Khani, and M. Ravansalar, “Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review,” *Chemometrics and Intelligent Laboratory Systems*, vol. 200, Article ID 103978, 2020.

[24] L. Chen, T. Wu, Z. Wang, X. Lin, and Y. Cai, “A novel hybrid BPNN model based on adaptive evolutionary Artificial Bee Colony algorithm for water quality index prediction,” *Ecological Indicators*, vol. 146, Article ID 109882, 2023.

[25] H. Wan, R. Xu, M. Zhang, Y. Cai, J. Li, and X. Shen, “A novel model for water quality prediction caused by non-point sources pollution based on deep learning and feature extraction methods,” *Journal of Hydrology*, vol. 612, Part A, Article ID 128081, 2022.