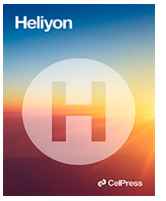
Heliyon 10 (2024) e37916

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Prediction of mine water quality by the Seq2Seq model based on attention mechanism

Xiaolong Wang [a], Yang Li [b,*]

[a] CHN Shendong Coal Group Co., LTD., Shenmu, 719300, China
[b] Summit Technologies Co., LTD, Xian, 710000, China

ARTICLE INFO

Keywords:
Water quality prediction
Artificial neural network
Sequence-to-Sequence model
Attention mechanism

ABSTRACT

In recent years, as China's industrialization level has advanced, the issue of environmental pollution, particularly mine water pollution, has become increasingly severe. Water quality prediction is a fundamental aspect of water resource protection and a critical approach to addressing the water resource crisis. For improvement in water quality prediction, this research first analyzes the characteristics of mine water quality changes and provides a brief overview of water quality prediction. Subsequently, the Long Short-Term Memory and Sequence to Sequence (Seq2Seq) models, derived from Artificial Neural Networks, are introduced. The Seq2Seq water quality prediction model is implemented, incorporating the attention mechanism. Experimental validation confirms the effectiveness of the proposed model. The results demonstrate that the attention mechanism-based Seq2Seq model accurately predicts parameters such as pH value, Dissolved Oxygen, ammonia nitrogen, and Chemical Oxygen Demand, exhibiting a high degree of consistency with actual results. They play a vital role in assessing the health of the water and its ability to support aquatic life. The change of these indicators can reflect the degree and type of water pollution. Moreover, the Seq2Seq + attention model stands out with the lowest predicted Root Mean Square Error of 0.309. Notably, in comparison to the traditional Seq2Seq model, the incorporation of attention mechanisms in the Seq2Seq model results in a substantial 2.94 reduction in Mean Absolute Error. This research on the Seq2Seq water quality prediction model with attention mechanism provides valuable insights and references for future endeavors in water quality prediction.

1. Introduction

With the growing global demand for natural resources, the exploitation of mineral resources has become an integral part. However, the impact of mining activities on the environment has aroused widespread concern, especially in the management of mine water. Mine water, as one of the main sources of wastewater produced in the mining process, is related to the health and sustainable development of the ecological environment in the mining area. Moreover, it also directly affects the production safety of miners and the effective utilization of mineral resources. Therefore, the monitoring and prediction of mine water quality is of great significance to guide the mine water resources' rational utilization, ensure the environmental safety of the mine, and promote the sustainable development of the mine economy.

* Corresponding author.
E-mail address: dly_1@163.com (Y. Li).

https://doi.org/10.1016/j.heliyon.2024.e37916
Received 14 January 2024; Received in revised form 20 June 2024; Accepted 12 September 2024
Available online 20 September 2024

In the context of current environmental protection and resource management, the demand for mine water quality prediction is increasingly urgent. The trend of water quality deterioration can be found in time by predicting mine water quality. Corresponding treatment measures can be taken to reduce the negative impact on the environment. At the same time, the demand for safe production and effective utilization of mine water resources are ensured. However, the prediction of mine water quality faces a series of challenges. Firstly, the water quality parameters of mine water are affected by many factors, including geological conditions of the mining area, mining method, depth and structure of mine, and variations in the surrounding environment, etc., which makes the change in mine water quality highly complex and dynamic. Secondly, traditional water quality monitoring and prediction methods often rely on regular water sample collection and laboratory analysis. It is not only time-consuming and labor-intensive but also difficult to achieve real-time monitoring and prediction of mine water quality changes. In addition, many traditional water quality prediction models are often unsatisfactory when dealing with nonlinear and non-stationary time series data, and it is difficult to accurately capture the complex model of mine water quality change [1].

With the rapid growth of society, the availability of large-scale historical data poses challenges in achieving accurate predictions. Effectively utilizing big data to forecast future system trends is a major problem. The emergence of deep learning (DL) provides a solution to this challenge by leveraging massive data to train DL models and obtain reliable predictions [2].

Hence, in the domain of mine water quality prediction, although the existing methods offer a variety of solutions, they have certain limitations. Past research has focused on using traditional statistical methods, machine learning techniques, and more recently the emerging DL method to predict water quality indicators. First, traditional statistical methods, such as autoregressive models, often assume that the data is linear and stable. This is difficult to meet in the actual water quality data processing because the water quality data usually presents nonlinear and non-stationary characteristics. Additionally, machine learning methods such as the artificial neural network (ANN) and Long Short-Term Memory (LSTM) exhibit advantages in processing nonlinear data. Nevertheless, they often require significant computational resources in training complex models and are susceptible to overfitting, which limits their application efficiency and generalization ability in water quality prediction. Meanwhile, the existing methods still need to be enhanced in the model's explanatory ability and the ability to predict future data. Consequently, finding a new method that can effectively deal with the nonlinearity of water quality data, reduce the risk of overfitting, and improve prediction accuracy and calculation efficiency has become a vital driving force and innovation point of this research.

With the advancement of DL, ANN possesses robust real-time data processing capabilities. However, when applying neural network models to water quality prediction, issues like convergence speed hinder the model's predictive performance. Consequently, this research aims to design a Seq2Seq water quality prediction model based on the attention mechanism. This approach enhances traditional Seq2Seq models by incorporating attention mechanisms and substantiates the proposed method's feasibility and effectiveness through experiments. The primary contribution of this research lies in the innovative utilization of the Seq2Seq model and attention mechanism to address mine water quality prediction challenges. By predicting continuous time series, the accuracy of water quality parameter predictions is improved. Dissolved Oxygen (DO), pH, ammonia nitrogen (NH_3-N), and Chemical Oxygen Demand (COD), are selected as the research objects, which are commonly used indicators in water quality assessment and are essential for monitoring the health of the water environment. The pH value has a direct impact on aquatic organisms. DO is the basic condition for the survival of aquatic organisms. NH_3-N and COD are important indicators to evaluate the degree of organic pollution. In addition, the prediction of these parameters can help to find pollution events in time and provide a scientific basis for water quality management and pollution control.

In comparison to other works, this research concentrates on continuous time series prediction, showcasing greater real-time and adaptive capabilities. Concurrently, the model's high accuracy across multiple crucial water quality parameters through experiments is validated, providing an effective solution to mine water pollution issues. Unlike other research, this research's focused subject matter underscores the significance and urgency of mine water quality concerns. This research bridges a research gap in the water quality prediction field related to this problem. Furthermore, the designed model's superior performance in water quality prediction is substantiated through comprehensive experimental validation and comparative analysis. These research outcomes hold importance for water resource preservation and mining environmental protection and offer valuable insights and references for the application of DL in time series analysis.

Building upon the above background, this research focuses on the variations in mine water quality and emphasizes the significance and content of water quality prediction. Leveraging the foundation of the ANN, this research delves into the LSTM model and the Seq2Seq model, specifically designing the water quality prediction model based on the Seq2Seq + attention (SSA). By incorporating the attention mechanism into the traditional Seq2Seq model, the feasibility and effectiveness of this approach are validated through experimental analysis.

The innovation encompasses several critical aspects. (1) By incorporating DL neural network models, this research effectively harnesses their inherent advantages in handling time series data. Compared to conventional prediction methods, DL models automatically extract features and patterns from data, thereby enhancing the accuracy of capturing trends in water quality parameter changes. (2) The integration of the attention mechanism enables the multi-step prediction of water quality parameters. This model's architecture capitalizes on the characteristics of time series data and adeptly addresses the complexity of multi-step forecasting, thus demonstrating considerable practical applicability. Integrating DL models into water quality prediction enhances predictive precision and operational efficiency, offering robust support for environmental preservation and water resource management. (3) In the analysis of the results, particular emphasis is placed on the comprehensive examination of the relationship between simulated and observed data, alongside a statistical exploration of the interdependencies among water quality parameters. This meticulous data analysis introduces novel methodologies and fresh perspectives to the realm of water quality prediction research.

2. Literature review

In the mine water quality prediction field, the application of DL models has become a research hotspot, and these models have been widely concerned because of their excellent performance in processing time series data. In this section, various methods of time series data prediction using the DL model are discussed, and their application and characteristics in mine water quality prediction are emphasized. These methods include the Sequence to Sequence (Seq2Seq) model, Recurrent Neural Network (RNN) and their variants, Convolutional Neural Network (CNN), and their comparison with traditional prediction methods. A careful review of the literature aims to reveal the potential of DL technology to improve the accuracy and efficiency of mine water quality prediction. First, the Seq2Seq model and attention mechanism-based methods are extensively applied in various time series prediction tasks, and their core advantages are more efficient processing of long series data and capturing time dependence. For example, Lim and Zohren (2021) [3] deeply analyzed the application of the Seq2Seq model in single and multi-step forecasting. However, Hao et al. (2023) [4] demonstrated the remarkable advantages of the Gated Recurrent Unit (GRU) model based on attention mechanism in improving prediction performance.

Then, RNNs and their variants, such as Long Short-Term Memory (LSTM) and GRU, become another vital class of models due to their ability to remember long-term dependent information when processing time series data. Gasparin, Lukovic, and Alippi (2022) [5] improved the application of RNNs in smart grid load prediction. Debow et al. (2023) [6] made a breakthrough in predicting the water quality index by using a stacked LSTM model. In addition, the CNN model provided a new perspective for complex prediction tasks by applying image conversion and multi-resolution imaging methods to time series data. The work of Barra et al. (2020) [7] was a typical example, in which they successfully applied CNN models to predict future market trends by converting time series data into Gramian Angular Field images.

Moreover, Song et al. (2024) [8] proposed the water resources carrying capacity risk index system and corresponding ranking criteria based on 20 influencing factors from four aspects: water resources endowment, economy, society, and ecological environment. Dong et al. (2023) [9] introduced a hybrid model based on signal decomposition and DL fusion to predict river water quality. Wang et al. (2023) [10] presented a short-term water quality prediction model by variational mode decomposition and an improved grasshopper optimization algorithm to optimize the LSTM neural network. These studies further validated the DL model's effective application in water quality detection.

Lastly, the DL model's performance was further verified by comparison with traditional prediction methods. Fu et al. (2023) [11] demonstrated the superiority of the Particle Swarm Optimation-Support Vector Regression (PSO-SVR) technique in water quality prediction compared with traditional neural network methods through comparison. However, Dritsas et al. (2023) [12] notably improved the prediction accuracy of the DL model by introducing synthetic minority oversampling technology. It can be found that the Seq2Seq model, especially when it incorporates an attention mechanism, shows dominant advantages in processing long time series data and improving predictive performance. This method can effectively capture long-term dependencies in time series, making it especially valuable in multi-step forecasting tasks.

Aslam et al. (2022) [13] collected water samples from wells in the study area (Northern Pakistan) to implement a river water quality index prediction model. They used four independent machine learning algorithms to assess water quality and help policy-makers in the China-Pakistan Economic Corridor region to better manage water resources. Liu et al. (2023) [14] proposed a novel prediction framework, which combined the two-stage feature selection model Golden Jackal optimization algorithm and the hybrid DL model. The purpose was to effectively capture the nonlinear relationship of multivariate time series in wastewater treatment plants. Talukdar et al. (2023) [15] introduced a DL-based stackable integrated model to predict the water quality index by integrating generalized linear models, gradient boosting machines, and neural network models. The uncertainty analysis denoted that the conductivity and total dissolved solids (TDS) had the highest uncertainty in predicting the water quality index.

It can be found that in the field of mine water quality prediction, the DL model's application has become a hot research spot, and these models have been widely concerned because of their excellent performance in processing time series data. The SSA method demonstrate significant advantages in predicting long-term dependence and multi-step prediction tasks. However, there are still several limitations that need to be addressed in the current study, including the dependence on massive high-quality data and the high computational complexity and resource requirements of the model. In addition, RNNs and their variants, such as LSTM and GRU, have become important tools in mine water quality prediction due to their ability to remember long-term dependent information. However, these models still face challenges regarding computing resource consumption and data complexity processing in real-time applications. For the CNN model, image conversion and multi-resolution imaging methods provide a new perspective for complex prediction tasks by converting time series data into images. Nevertheless, these methods still have shortcomings in data preprocessing and model interpretability.

The proposed method has advantages and improvements in many aspects. Firstly, the prediction accuracy and stability of long-term change in mine water quality can be improved by combining the Seq2Seq model's time-dependent advantage with the attention mechanism's accurate capturing ability. Secondly, through optimized data enhancement and preprocessing techniques, the model's robustness to noise and outliers is effectively enhanced, thus improving the reliability of the prediction results. Finally, advanced model evaluation and optimization techniques, such as cross-validation and hyperparameter adjustment, ensure the model's generalization ability and performance stability across different scenarios and datasets. In summary, although the DL model has shown significant potential and advantages in mine water quality prediction, it still needs to be further improved and optimized in terms of data acquisition, computational resources, and interpretability. By integrating cutting-edge technologies and approaches, the proposed method aims to overcome the limitations of current methods and offer new exploration directions and implementation strategies for future research.

3. Materials and methods

3.1. Variation characteristics and prediction of mine water quality

The mine water, influenced by human activities during the coal mining process, is characterized by its close contact with coal and rock layers, leading to a series of physical and chemical reactions. Consequently, mine water exhibits distinct characteristics associated with the coal industry [16]. For instance, it typically possesses poor sensory properties and exceeds standard limits for suspended solids. Additionally, mine water often bears surface oil stains and contains waste engine oil, emulsified oil, and other organic substances [17]. As a result, the water environment becomes relatively complex, and the corresponding water quality data assumes particular traits: complexity, periodicity, and relevance [18].

1. Complexity: The water environment represents an intricate system wherein water quality changes are susceptible to human activities and hydrogeological structures. These factors contribute to variations in the physicochemical parameters of water, introducing a degree of uncertainty.
2. Periodicity: Water temperature significantly influences chemical and physical indicators. During winter, lower temperatures reduce microbial activity, affecting the decomposition of organic matter and leading to lower levels of DO. Conversely, summer exhibits the opposite effect. These indicators display periodicity due to the cyclic changes in water temperature.
3. Relevance: Alterations in water indicators often trigger changes in other indicators. For example, there is a noticeable autocorrelation between the concentrations of NH_3-N and DO. Similarly, the concentration of DO shows a positive correlation with chlorophyll. Thus, there exists a correlation between various physical and chemical indices of water quality.

Groundwater quality assessment typically relies on various physicochemical indexes, including water temperature, pH value, conductivity, TDS, DO, COD, sulfate (SO_4^2), NH_3-N, Total Phosphorus, sodium (Na), as well as certain heavy metals like mercury (Hg), arsenic (As), iron (Fe), copper (Cu), manganese (Mn), lead (Pb), cadmium (Cd), chromium (Cr), among others [19].

In China, the Ministry of Environmental Protection has established the Quality Standards for Groundwater (GB/T 14848-2017) [20]. These standards provide a framework for assessing groundwater quality and ensuring its suitability for different purposes. Authorities can effectively manage and protect groundwater resources by monitoring and evaluating these physicochemical indexes.

In light of these criteria, water quality prediction processes involve forecasting the future changes in the quality of a specific water body using mathematical models based on measured water quality data [21]. This prediction encompasses two approaches: point source pollution and non-point source pollution. Point source pollution prediction includes estimating the pollution levels in water bodies, assessing the water quality's capacity to accommodate sewage and other pollutants, forecasting potential pollution scenarios resulting from the construction and operation of proposed plants, and predicting improvements in water quality following the implementation of pollution prevention measures. Accurate water quality prediction enables managers to monitor and regulate water quality fluctuations effectively [22].

Time series data plays a crucial role in the field of water quality prediction. Time series refers to a collection of data points obtained through continuous monitoring of one or more system variables over a specific period [23]. These data points are arranged in chronological order, forming a sequence that can be utilized to predict future events or outcomes. DL leverages the power of big data to train models and generate accurate predictions [24]. In the context of water quality, many physicochemical indices are recorded as time series data, organized based on the monitoring time sequence. Therefore, employing time series prediction methods becomes instrumental in forecasting water quality changes [25]. Advancements in science and technology have led to a decrease in the cost of water quality monitoring sensors. Consequently, a multitude of sensors can be deployed around water bodies, providing real-time data that can be utilized for DL-based correlation algorithms. By training on this continuously updated data, DL models can effectively predict future trends and variations in water quality, enabling timely interventions and proactive management.

3.2. ANN and LSTM models

Water quality changes are inherently nonlinear, making it challenging for traditional linear regression models to address such complexities effectively. ANN offers a nonlinear modeling approach by utilizing a large number of artificial neurons to estimate functions based on statistical principles [26]. ANN consists of input, output, and hidden layers, with each layer's output serving as the input for the subsequent layer [27]. To accurately predict the time series of water quality, a single hidden layer with a single input node as the prediction node is designed $(l \times k \times 1)$, as denoted in Eq. (1).

$$y_{t'} = \sum_{m=1}^{k} \varepsilon_m w \left(\sum_{n=1}^{l} \varepsilon_{nm} x_t + u_m \right) + u_0 \tag{1}$$

y means the output of the NN; x refers to the input; ε stands for the weight that the NN can learn; w reveals the activation function; u shows the bias value.

LSTM, a type of RNN variant, is commonly employed for time series prediction with long prediction intervals and delays [28]. The unit structure of LSTM networks is illustrated in Fig. 1, showcasing its ability to capture long-term dependencies in time series data.

Fig. 1 depicts an LSTM unit, which comprises a memory unit and three gate units: the input, output, and forget gates. These gates

play a crucial role in the information update process of LSTM at time t [29].

$$a_t = \varphi[Q_a(f_{t-1}, x_t) + h_a] \tag{2}$$

$$b_t = \varphi[Q_b(f_{t-1}, x_t) + h_b] \tag{3}$$

$$\widetilde{c}_t = \tanh\,[Q_c(f_{t-1}, x_t) + h_c] \tag{4}$$

$$c_t = a_t \odot c_{t-1} + b_t \odot \widetilde{c}_t \tag{5}$$

$$d_t = \varphi[Q_d(f_{t-1}, x_t) + h_d] \tag{6}$$

$$f_t = d_t \odot tan\, hc_t \tag{7}$$

a_t, b_t, and c_t signify the input gate, forgetting gate, and output gate, respectively; d_t expresses the memory unit; f_t stands for the hidden state variable; φ represents the sigmoid activation function; tanh means the tanh function; Q stands for the parameter matrix of gate and memory unit; x implies input value; h shows bias. The LSTM model, known for its memory gate and forgetting gate mechanisms, effectively handles the dependencies between features with substantial input intervals. These gates play a crucial role in updating the information state of the storage unit at a given time.

3.3. SSA

The Seq2Seq model is a NN characterized by a coding-decoding structure, where both the input and output are sequences [30]. The encoder component transforms a time sequence of variable length into a fixed-length vector, while the decoder component converts a fixed-length vector into a target time sequence of variable length [31]. The structure of the Seq2Seq model is illustrated in Fig. 2.

Fig. 2 illustrates that the Seq2Seq model utilizes LSTM as the encoding-decoding unit. It sequentially encodes the input sequences of length L, with each input being processed individually. The output memory unit c_L represents the feature information extracted from the input sequence. The state vector $s_{t'}$ is generated based on the decoder's output information and is calculated using c_L and the state vector $s_{t'+1}$ from the previous step. The state vector $s_{t'+1}$ is updated during the decoder's prediction of the next time series. The calculation of the state vector at time $t' + 1$ is expressed as Eq. (8) [32].

$$s_{t'+1} = a_1(s_{t'}, c_L) \tag{8}$$

a_1 is the input gate in LSTM. Through extensive training, LSTM can extract valuable feature information from complex time series data.

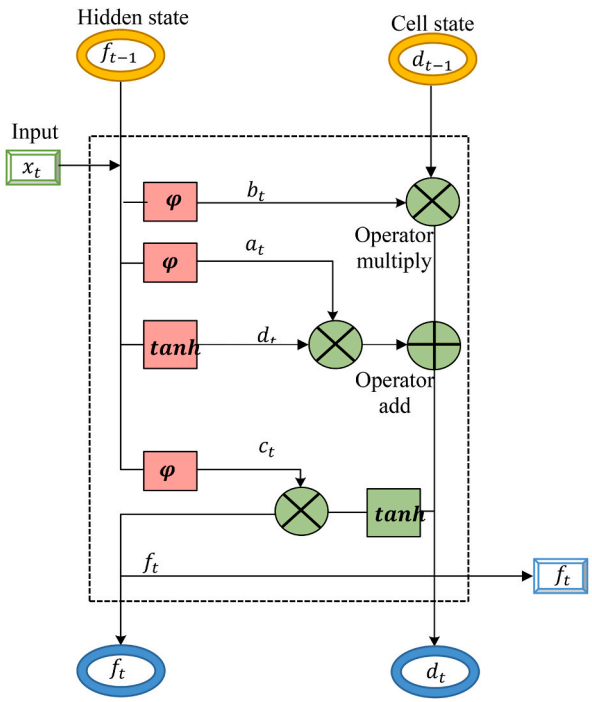


Fig. 1. Unit structure of the LSTM model.

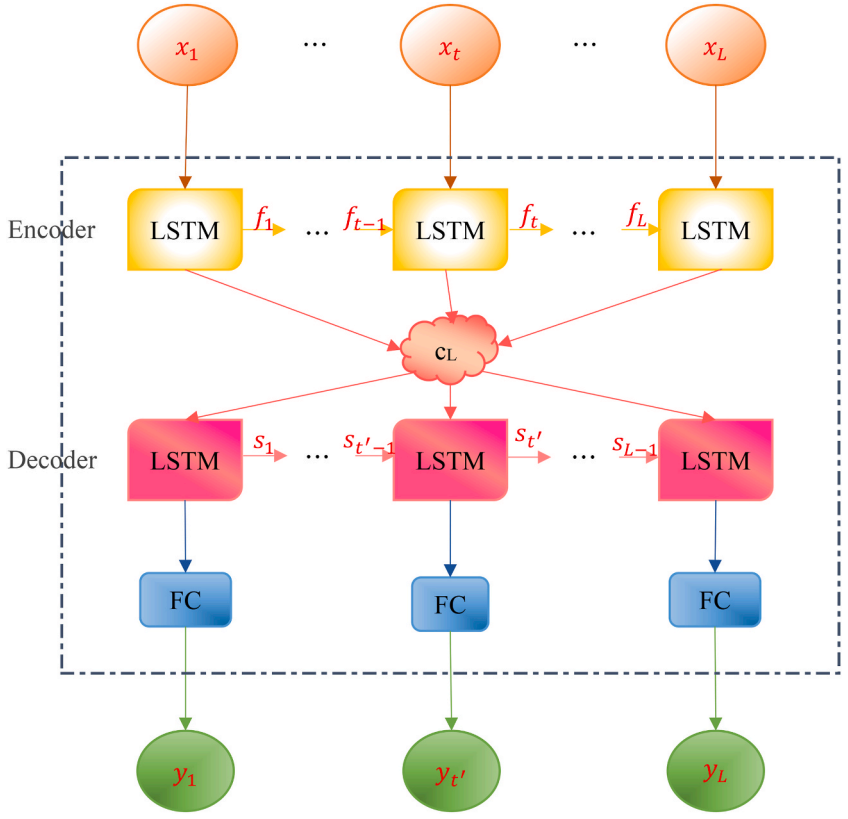


Fig. 2. Structure of Seq2Seq model.

The final Fully Connected (FC) layer in the model is responsible for decoding this extracted information into predictive values, as shown in Eq. (9).

$$\hat{y} = \alpha_k^L s \tag{9}$$

$\hat{y}$ means the forecast value generated by the model; α represents the weight of the FC layer; s refers to the state vector extracted by LSTM.

In the Seq2Seq model, there is a potential loss of information during the encoding and decoding process, which limits the ability of the encoder to effectively focus on the finer details of the input sequence [33]. The attention mechanism is introduced to address this issue, which encodes the sequence into multiple memory units using different time steps. The decoder then utilizes these memory units to generate more accurate output results [34]. The structure of SSA is presented in Fig. 3.

In Fig. 3, the attention layer incorporated into the Seq2Seq model enables independent selection of the state vectors from all time steps T generated by the corresponding encoder [35]. The attention weight for each encoder's state vector is determined by the decoder's previous state vector $s_{t'-1}$ and the LSTM unit's state vector s_T'. Eq. (10) details the calculation of attention weight γ [36].

$$\gamma_{t'}^n = U_s^T \tanh\left[Q_s\left(s_{t'-1}, s_T'\right) + V_s s_n\right], 1 \le n \le T \tag{10}$$

s_n represents the state vector of the encoder at time n, while U_s and Q_s denote the parameters that V_s needs to learn from the model. Eq. (11) demonstrates the weight calculation of the state vector of the nth encoder in relation to the prediction value at time t'.

$$\delta_{t'}^m = \frac{exp\ \gamma_{t'}^n}{\sum_{n=1}^{T} exp\ \gamma_{t'}^n} \tag{11}$$

$\delta_{t'}^m$ is used to obtain memory unit $d_{t'}$,which is calculated according to Eq. (12).

$$d_{t'} = \sum_{m=1}^{T} \delta_{t'}^m s_m \tag{12}$$

After obtaining the memory unit summed by weights, it can be converted into the input of the decoder, as plotted in Eq. (13).

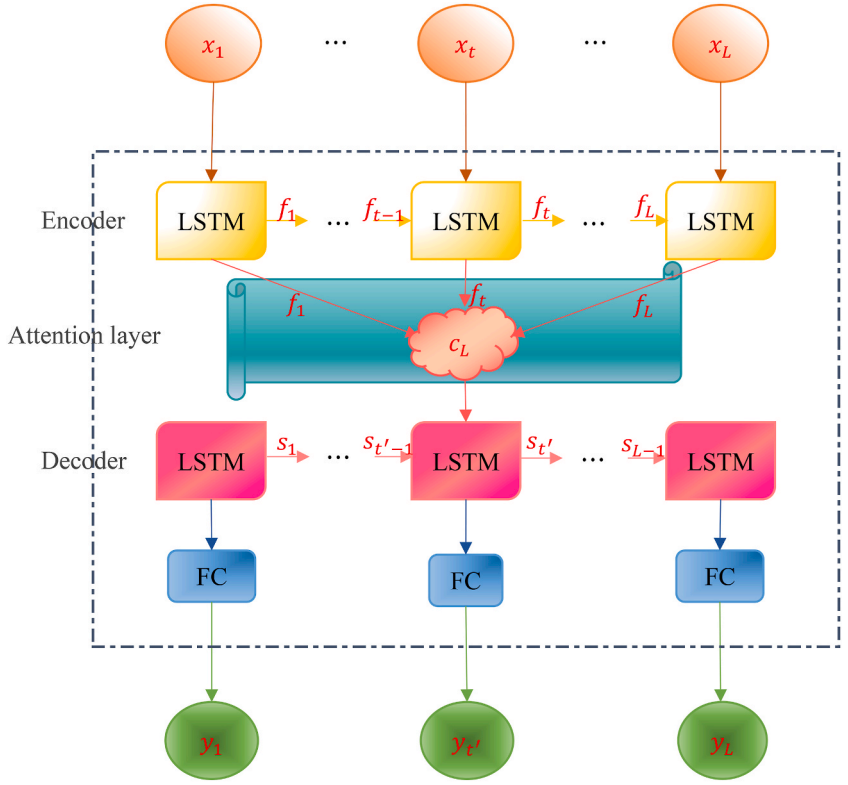


Fig. 3. Water quality prediction model of SSA.

$$\widetilde{y_{t'-1}} = \widetilde{\alpha^T} d_{t'-1} + \tilde{a} \tag{13}$$

$d_{t'-1}$ signifies the memory unit calculated by the encoder in the previous step; $\widetilde{\alpha^T}$ and $\tilde{a}$ express a parameter that requires model learning. $\widetilde{y_{t'-1}}$ is employed to update the decoded state vector at time t'. Eq. (14) describes the input of the decoder.

$$s_{t'} = a_2(s_{t'-1}, \widetilde{y_{t'-1}}) \tag{14}$$

a_2 stands for an input gate of the LSTM unit. The final output value at time t' reads:

$$\widehat{y_{t'}} = D_y s_{t'} + a_y \tag{15}$$

In Eq. (15), $\widehat{y_{t'}}$ refers to the result of the output value at time t'; $s_{t'}$ means the state vector of the decoder; D_y and a_y stand for the parameter of the linear model.

3.4. Experimental preparation

The water quality prediction model introduced here is implemented and trained using the Keras 2.1.3 DL framework. The experimental configuration includes the utilization of the Windows 10 (64-bit) operating system, equipped with an Intel(R) Core(TM) i5-9500 CPU and an NVIDIA GeForce GTX1080Ti GPU. The Anaconda 3 environment management tool is employed to facilitate the experiment. The employed water quality dataset encompasses the physical and chemical attributes of mine water quality within area A. Mining area A is located in the west of Liaoning Province, which is one of the important iron ore mining bases in China. The geographical location of the area is about 41° north latitude and 122° east longitude, and it belongs to the temperate monsoon climate. The selection of the mine area considers its unique geographical environment and mineral resources, as well as Liaoning Province's position as a major industrial and mineral mining town in China. Besides, the water quality in the area has been significantly affected by mineral mining activities and surrounding industrial activities, making it an ideal location to study changes in mine water quality. The water quality in this area is greatly affected by specific mining activities, which makes the study of the water quality in this area have unique research value and practical significance. Recognizing the remarkable differences in water quality and environment between North and South China, which are mainly caused by factors such as geological structure, climatic conditions, and human activities. The northern regions generally face water shortages, while the south suffers from environmental problems such as acid rain. Hence, mining area A is selected as the research object to explore the water quality prediction model for areas with specific

geographical and environmental characteristics.

Spanning the timeframe from 2016 to 2021, this dataset is sourced from the national groundwater information service platform. Each entry in the dataset encompasses 16 physicochemical indicators, including parameters such as transparency, water depth, water temperature, conductivity, suspended solids, DO, pH, silicate, phosphate, nitrite nitrogen, nitrate nitrogen, ammonia nitrogen, permanganate index, total phosphorus, total nitrogen, and dissolved total organic carbon. To train and verify the proposed model more effectively, the dataset is divided into three parts: training, verification, and test sets. The specific allocation ratio is 70 % for the training set, 15 % for the verification set, and 15 % for the test set. The purpose of this segmentation ratio is to ensure that the model can be trained on a sufficient amount of data while leaving a certain amount of data for model performance verification and testing, to avoid over-fitting and improve the model's generalization ability. Four water quality monitoring indicators, namely pH value, DO, NH_3-N, and COD, are selected for analysis.

3.5. Selection of time series characteristics and evaluation indexes for water quality prediction

It is crucial to consider the various water quality indicators' diverse effects to enhance the accuracy of water quality prediction. The model's performance is analyzed by the combination of descriptive statistics and inferential statistics. In this research, the Spearman correlation coefficient is employed to assess the correlation between time series data and identify features with high correlation for training purposes. This is mainly because it can capture nonlinear relationships, and it is also very robust for non-normally distributed data. By identifying features in time series data that are highly correlated with target variables, it is possible to ensure that model training is focused on the most influential factors, thereby improving the accuracy and efficiency of predictions. Spearman correlation coefficients are particularly suitable for processing environmental and water quality data where nonlinear relationships may exist. The calculation of the Spearman correlation coefficient (ϑ) is presented in Eq. (16) [37].

$$\vartheta = 1 - \frac{6\sum_{t=1}^{N}(X_t - Y_t)}{n(n^2-1)} \tag{16}$$

N denotes the sample size; X_t and Y_t refer to the values of time series X and Y at time t.

Three evaluation metrics are utilized to assess the performance of water quality prediction, namely, root mean square error (RMSE), mean absolute error (MAE), and determination coefficient R^2. In water quality prediction, large prediction errors can lead to incorrect water quality management decisions, so using RMSE can guarantee that the model pays sufficient attention to these large errors. In the context of water quality prediction, a high R^2 value indicates that the model can accurately capture the dynamic characteristics of water quality parameters over time, thus affording reliable information for water quality management. Through the comprehensive use of these evaluation indexes, the water quality prediction model's performance can be comprehensively evaluated from diverse angles to ensure that the developed model has high accuracy and practical value. When the RMSE value is lower than the mean of the prediction target, it indicates that the model possesses a minor prediction error and exhibits proficient predictive capability. RMSE is a widely used index for evaluating measurement accuracy, and its calculation is outlined in Eq. (17).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2} \tag{17}$$

The smaller the RMSE value, the more accurate the prediction result. MAE measures the error of the model in prediction. A smaller MAE corresponds to a reduced prediction error of the model, indicative of its closer proximity to the actual observations on average. Its calculation is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i - y_i'\right)\right| \tag{18}$$

The smaller the MAE value, the better the prediction result. The determination coefficient (R^2) is used to reflect the prediction model's fitting degree. The R^2 value represents the proportion of the variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, where a value closer to 1 indicates a better-fitting effect. As the R^2 value approaches 1, it indicates that the model can explain a larger proportion of the variability in the dependent variable, resulting in a more accurate and reliable prediction. A higher R^2 value, approaching 1, signifies a stronger alignment between the model's predictions and the actual observed values, illustrating the model's efficacy in elucidating data variability. Conversely, an R^2 value closer to 0 indicates a lesser proportion of variance accounted for by the model, reflecting poorer predictive performance. R^2 is calculated via Eq. (19).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left|y_i - y_i'\right|}{\sum_{i=1}^{n}\left|y_i - \bar{y}\right|} \tag{19}$$

In Eqs. (17) and (18), Eq. (19), y_i and y_i' represent the real value and predicted value of the monitoring data at time I; $\bar{y}$ means the average value of real-time monitoring data; n signifies the total capacity of sequence data.

4. Results and discussion

4.1. Data distribution of water quality DO and NH_3-N

Fig. 4 displays the historical data distribution for DO and NH_3-N as examples.

Fig. 4 illustrates the distribution trend of two water quality indicators, namely DO and NH_3-N, in mine water in zone *A* from 2016 to 2021. The DO values range from 5 to 13 mg/L and exhibit noticeable periodicity. Besides, the NH_3-N values range from 0 to 3.5 mg/L, with the majority falling within the range of 0–1 mg/L. In exploring changes in mine water quality, this research focuses specifically on how seasonal factors affect water quality parameters, including pH, DO, NH_3-N, and COD. Seasonal changes have prominent effects on water temperature, precipitation, and biological activities, which further affect water quality.

In spring and autumn, the DO value is higher, which may be related to lower water temperature and higher gas solubility. In summer, the DO value decreases, which may be due to higher temperatures and increased microbial activity, leading to increased oxygen consumption. In winter, DO values increase again, reflecting lower levels of biological activity at low temperatures. The value of NH_3-N peaks in summer, possibly due to high temperatures and accelerated decomposition of organic matter, resulting in increased release of NH_3-N. In winter, the value of NH_3-N is lower, which may be related to the decrease in microbial activity. In addition, through the comparative analysis of water quality parameters in different seasons, it is found that the water quality in spring and autumn is relatively good. However, due to high temperatures and vigorous biological activities, water quality parameters such as NH_3-N and COD are often higher in summer. These findings have vital implications for the management and treatment of mine water. These can help formulate water quality management strategies for diverse seasons, and improve the effectiveness of water quality monitoring and pollution control.

To understand this phenomenon, the variations of water quality parameters in spring, summer, autumn, and winter are analyzed comprehensively. Table 1 summarizes the variation range of water quality parameters in each season, and reveals the regularity and characteristics of seasonal factors' influence on mine water quality through comparative analysis. This analysis not only helps to better understand the seasonal variation of water quality in mining areas but also provides a scientific basis for formulating more effective water quality management strategies.

Table 1 illustrates the obvious change trend of water quality parameters in various seasons. In spring, the DO level increases with the rise of rain. In summer, the increase in water temperature leads to an increase in COD value. The water quality in autumn is relatively stable, while in winter, due to the influence of low temperature, the DO level adds and the COD value decreases. These changes show the influence of seasonal factors on water quality parameters. It can be found that the influence of seasonal factors on mine water quality cannot be ignored.

The seasonal analysis of this research reveals the regularity of water quality parameters changing with seasons, which is of great significance for the formulation of targeted water quality management measures. For instance, the increase of COD caused by the increase in water temperature in summer suggests the need to strengthen the treatment of organic pollutants in this season. The increase of DO levels in winter provides more favorable conditions for biological treatment. Through an in-depth understanding of seasonal changes, the mine water quality can be predicted and managed more effectively, and the scientific basis for environmental protection in mining areas can be offered. Through the in-depth analysis of water quality parameters in different seasons, it not only reveals the significant impact of seasonal changes on mine water quality but also provides a valuable reference for future water quality management and treatment strategies. These findings help to better cope with the challenges of seasonal changes to water quality in mining areas and ensure more scientific and efficient water quality management.

The descriptive statistical analysis results for water quality indicators are depicted in Fig. 5. It illustrates that distinct variations exist in the distribution characteristics of different water quality indicators. The pH distribution of the water quality data shows a

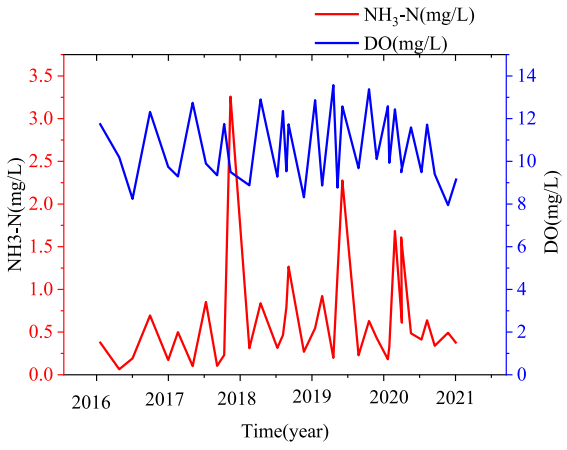


Fig. 4. Historical data distribution of water quality DO and $NH_3 - N$.

Table 1
Effect of seasonal variation on water quality parameters.

Season	Range of ph	DO (mg/L)	NH_3-N (mg/L)	COD (mg/L)	Annotation
Spring	6.5–7.5	8–10	0.5–0.8	60–90	With more rain, DO levels rise
Summer	6.8–8.0	6–9	0.6–1.0	70–100	As water temperature increases, COD increases
Autumn,	7.0–7.8	9–11	0.4–0.7	50–80	Water quality is relatively stable
Winter	6.7–7.6	10–12	0.3–0.5	40–70	Low temperature, high DO level, COD reduction

relatively concentrated pattern, with an average value closely aligned with neutrality. It depicts that the pH of mine water changes little and is maintained in the near neutral range. This is of positive significance to the ecological balance of the water body and mine water treatment. The mean pH value is about 7.2 with a small standard deviation, indicating that the water quality is stable during the sampling period. The distribution of DO content is wider, accompanied by a larger standard deviation, suggesting substantial fluctuations. The change in the DO level may be seasonal, rainfall, and the influence on biological activity of mine. This volatility requires special attention because DO levels directly affect the living environment of aquatic organisms. Furthermore, upon examining the skewness, it becomes apparent that the distributions of NH_3-N and COD exhibit right-skewness. The distribution of NH_3-N shows a clear right skew, indicating that most samples have low NH_3-N content, but there are some high outliers. These outliers may result from specific contamination events during certain periods or from occasional failures of mine water treatment systems. The distribution of COD is also skewed to the right, illustrating that in most cases, the content of organic matter in the water quality is low, but some sampling sites show higher COD values. This may be due to the concentration of organic pollutants discharged from mines at certain points in time or under certain conditions.

Here, to ensure the accuracy of the time series analysis, the Augmented Dickey-Fuller (ADF) test is used to evaluate the stationarity of the collected water quality data series. Stationarity is a fundamental premise in time series analysis, which guarantees that the model's statistical properties (such as mean and variance) remain constant over the entire series, which is crucial for subsequent predictive analysis. When conducting ADF tests, the main indicator of concern is the p-value, which is a statistical measure of the probability that the observed data can occur if the null hypothesis holds. In this research, the null hypothesis assumes that the sequence has a unit root, that is, the sequence is non-stationary. If the obtained p-value is less than the predetermined significance level (set at 0.05), there is sufficient evidence to reject the null hypothesis and consider the series to be stationary. This means that the statistical properties of these sequences remain unchanged throughout the observation period, meeting the requirements of time series analysis. Through the descriptive statistical analysis and stationarity test of water quality indicators, this research not only reveals the distribution characteristics and change rules of each indicator but also provides a solid foundation for the subsequent time series prediction analysis. Understanding these statistical characteristics is important for formulating effective water quality management strategies and optimizing mine water treatment processes.

In Fig. 6, through the ADF test of each water quality index sequence, it is found that the P-value of all sequences is significantly lower than the significance threshold of 0.05. This result shows that the individual sequences are statistically stationary, i.e. their statistical properties (such as mean and variance) are constant throughout the observation period. This indicates that the mean and variance of pH do not change significantly throughout the observation period, and the water quality is stable in terms of pH. Although DO values have some volatility, their statistical properties remain consistent across time series, providing a basis for the direct application of prediction models. NH_3-N sequence: P value is much lower than 0.05, and the sequence is stable. This illustrates that although there are outliers, the NH_3-N content is stable in the overall trend and can be effectively predicted in time series. The stationarity of COD shows that the organic pollutant content of COD is stable in time, which contributes to the prediction model's accuracy. This is a critical finding because it means that time series prediction models can be applied directly without additional stationarity processing, such as difference or transformation. These are often necessary steps for processing non-stationary sequences.

The Spearman correlation coefficient is calculated to analyze the correlation between these water quality indicators, and the results are exhibited in Table 2.

Table 2 reveals that the correlations between pH and DO, as well as pH and COD, exhibit relatively weak coefficients of 0.16 and 0.11, respectively. However, a notably stronger negative correlation is observed between pH and NH_3-N, with a coefficient of −0.52. This implies a potential inverse relationship between pH and NH_3-N content, suggesting that a decrease in pH could correspond to an

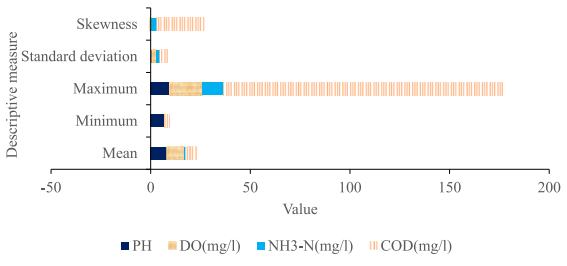


Fig. 5. Descriptive statistical analysis results of water quality indicators.

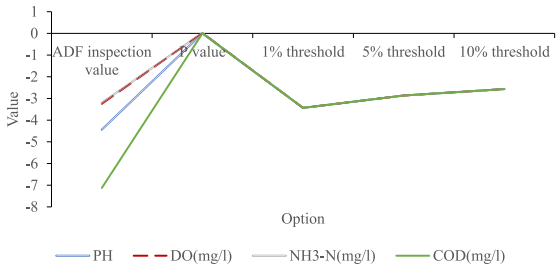


Fig. 6. Analysis results of stationarity of water quality series data.

increase in NH_3-N concentration. Comparatively, the correlation between DO and pH is similarly weak (0.16), as is the correlation between DO and NH_3-N (−0.05). In contrast, the correlation between DO and COD is slightly stronger, with a coefficient of −0.11. This observation may indicate that lower dissolved oxygen levels could be associated with higher COD concentrations within the context of water quality. The pronounced negative correlation (−0.52) between NH_3-N and pH suggests the possibility of an inverse relationship between ammonia nitrogen content and pH. In contrast, the correlation between NH_3-N and other indicators is comparatively weaker, showing a correlation of −0.05 with DO and 0.23 with COD. This implies a potential link between NH_3-N content and water acidity and alkalinity, possibly indicating a positive correlation with COD content. While the correlation between COD and pH is weak (0.11), and its relationship with DO is similarly weak (−0.11), a slightly stronger correlation emerges between COD and NH_3-N, characterized by a coefficient of 0.23. These findings imply that COD content may not be strongly linked to water acidity and alkalinity; however, a moderate positive correlation with NH_3-N content is feasible. The Spearman correlation coefficients among the water quality indicators demonstrate a significant correlation, as the absolute values of most coefficients exceed 0.1.

4.2. Test results of water quality prediction of SSA

The experiment conducts tests using different input step sizes to achieve the optimal prediction effect. It mainly focuses on the effect of comprehensive prediction of mine water quality by using the SSA model. This part of the test aims to evaluate the model's optimal prediction effect through the size of diverse input steps, and then compare the performance of different DL models in multi-step prediction. Key water quality indicators encompass pH, DO, TDS, and specific heavy metals. The corresponding results are illustrated in Fig. 7.

The analysis in Fig. 7 shows that the attention model's performance for RMSE and MAE reaches its maximum when the input step is set to 10. This means that at smaller input steps, the model exhibits a larger error in predicting water quality. It indicates that the model's ability to capture sequence data is limited, possibly because the shorter step size does not adequately capture the long-term dependence of the time series. In contrast, when the input step size is adjusted to 40, the minimum values of MAE and RMSE, and the maximum values of the determination coefficient R^2, are observed. This indicates that the SSA method used for water quality prediction produces the most stable and effective results at longer input steps. Longer input steps allow the model to capture longer-term dependencies and patterns in the time series data, improving the accuracy of predictions. This finding highlights the importance of choosing the appropriate input step size when implementing a time series prediction model. Too short a step size may not adequately capture the time dependence of the data, resulting in poor prediction accuracy. Appropriately increasing the step size enables the model to learn based on more comprehensive historical information, which can markedly enhance the stability and accuracy of the prediction.

In addition, this result highlights the SSA method's effectiveness when dealing with water quality prediction problems. By carefully selecting input steps, the SSA method optimizes model performance, minimizes prediction errors, and improves the model's ability to predict future water quality changes. This has vital practical implications for water quality management and decision-making processes, helping to identify and respond to water quality changes in advance, and ensuring the safe and sustainable use of water resources.

The prediction errors associated with distinct DL models across varying input step sizes are depicted in Fig. 8. Notably, an increase in the prediction step size corresponds to an elevation in the RMSE of each model. This pattern aligns with intuitive expectations, as extending the forecasting horizon introduces heightened uncertainty, consequently yielding larger errors. This trend highlights the intrinsic complexity of time series prediction. Amidst diverse input step sizes, the SSA model consistently exhibits superior

Table 2
Correlation analysis of water quality data.

	pH value	DO(mg/L)	NH_3-N (mg/L)	COD(mg/L)
pH value	1	0.16	−0.52	0.11
DO(mg/L)	0.16	1	−0.05	−0.11
NH_3-N (mg/L)	−0.52	−0.05	1	0.23
COD(mg/L)	0.11	−0.11	0.23	1

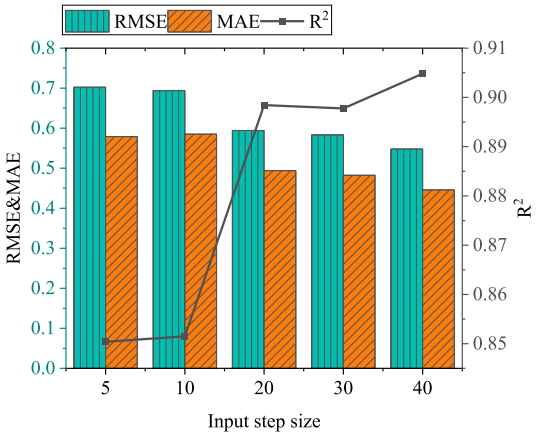


Fig. 7. Test results of SSA with different input step sizes.

performance in multi-step forecasting when compared to alternative models. Specifically, as input step sizes enlarge, this model consistently demonstrates relatively diminished prediction errors while other models grapple with persistently larger errors. This underscored enhancement underscores the efficacy of the attention mechanism in addressing the complexities of multi-step time series forecasting, effectively capturing essential sequence information. Contrasted against this, conventional time series forecasting techniques like Autoregressive Integrated Moving Average (ARIMA) and SVR display comparatively inferior performance in multi-step forecasting scenarios. Conversely, models such as Seq2Seq, SSA, ANN, and LSTM yield more reliable and precise outcomes for multi-step forecasting. Significantly, the SSA model consistently maintains minimal prediction errors across varying input step sizes, unequivocally affirming its prowess in confronting the intricacies of multi-step forecasting challenges.

4.3. Water quality prediction results of SSA

The water quality data of mine water in area A are divided into training and test sets in a ratio of 3:1. The SSA method is applied, and the prediction result for the pH value is presented in Fig. 9.

Fig. 9 details that the predicted pH values are relatively close to the actual results, with most of the data concentrated in the range of 7–8.5. However, there are occasional instances where the predicted pH values deviate from the actual values, resulting in either too-high or too-low values. The maximum predicted pH value is 9.1, while the minimum is 7.2. In comparison, the actual maximum and minimum pH values are 8.7 and 7.3. The difference between the two minimum values is minimal, but there is a difference of 0.4 in the maximum values. Overall, the disparity between the predicted and actual pH values falls within an acceptable range, indicating that the model is effective. Deviation sources: On the one hand, data noise: Mine water quality data may contain certain noise and outliers, which may affect the prediction results. On the other hand, there are model limitations: Although the SSA method has advantages in processing time series data, its predictive power is still limited by model complexity and data characteristics. Most of the deviations between the predicted and actual values are between 0.1 and 0.4, an acceptable error range in water quality monitoring. Most of the predicted values are close to the actual values, indicating that the model has a high accuracy in capturing the trend of water quality change. The SSA method's validity provides a reference for its application in other water quality index predictions, which is helpful to comprehensively improve the ability of mine water quality prediction.

The prediction results for DO are suggested in Fig. 10.

Fig. 10 demonstrates that the predicted results of DO exhibit relatively small deviations from the actual values, although there are

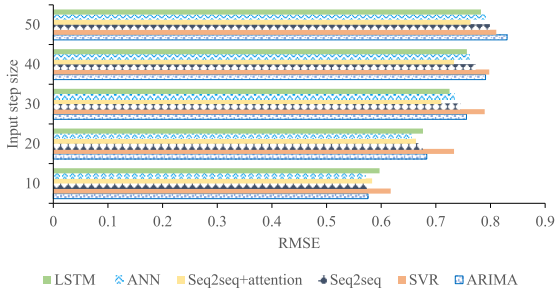


Fig. 8. Prediction errors of different DL models with different input step sizes.

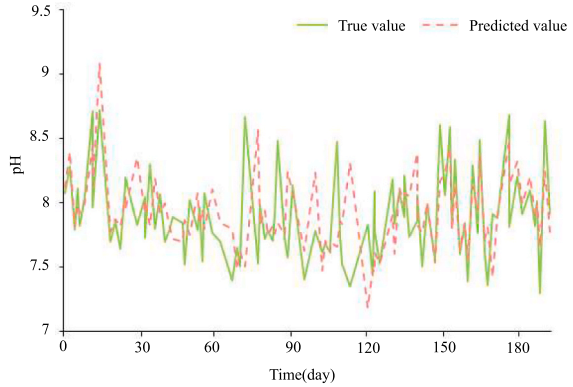


Fig. 9. Prediction results of pH value.

instances where the predictions differ significantly. The maximum difference between the predicted and actual DO values is 0.023 mg/L, while the minimum difference is 0.0004 mg/L. Notably, in the time interval of 90–120, the trend lines of the predicted and actual DO values coincide, indicating a periodic pattern. This shows that the model can capture the periodic change of DO value and has a good prediction effect. The majority of the predicted and actual DO values are concentrated in the range of 6–12 mg/L, further illustrating the designed model's effectiveness. Although the overall deviation is small, there are still significant deviations in some periods, which may be caused by the following factors. Sudden changes in the mine environment may cause abnormal fluctuations in the DO value. Errors and data noise during measurement can also affect the prediction results. However, the deviation range between the predicted and actual values is very small, and the maximum difference is only 0.023 mg/L, which is an acceptable error range in practical applications. In a specific time interval, the high agreement of trend line between the predicted and actual values indicates that the model can effectively identify and predict the periodic change of DO value. These results validate the model's validity and provide scientific support for real-time monitoring and management of mine water quality. By applying this prediction model, more accurate water quality monitoring and management can be realized to ensure the sustainable development of the mine environment. Fig. 11 showcases the prediction results for NH_3-N.

Fig. 11 illustrates that the highest predicted and actual values of NH_3-N are 3 mg/L and 3.5 mg/L, respectively. The minimum predicted value of NH_3-N is 0.2 mg/L, while the actual value is 0.17 mg/L. The difference between the highest and minimum values is 0.5 mg/L and 0.03 mg/L, respectively, indicating a relatively small variation. In general, the predicted NH_3- NH_3-N values closely align with the actual results, and both are concentrated within the range of 0–1.5 mg/L, affirming the model's effectiveness. Source of error: Deviation between predicted and actual values may be due to errors in the measurement process and external interference. In addition, although the difference between the highest and lowest values exists, these differences are within the acceptable range in the practical application of environmental monitoring. The prediction results for COD are destroyed in Fig. 12.

In Fig. 12, there is a significant disparity between the predicted and actual results of COD. The largest gap occurs when the actual COD value reaches its maximum, with a difference of 3.3 mg/L. However, overall, the difference between the predicted and actual results is relatively small, concentrated within the range of 3.0–7.0 mg/L. This gap denotes that the SSA model developed here demonstrates satisfactory performance in water quality prediction. Although there are some extreme value differences, these

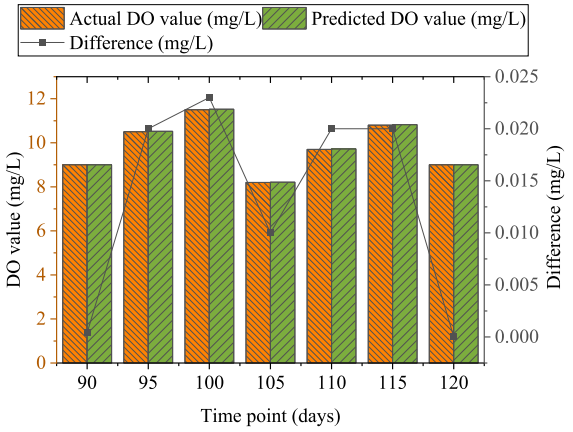


Fig. 10. Prediction result of DO.

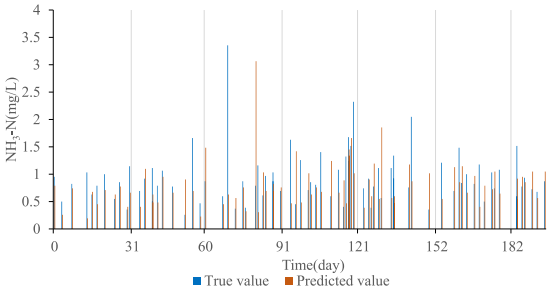


Fig. 11. Prediction results of NH_3-N.

differences are generally acceptable in environmental monitoring. Generally, the model can provide valid predictions.

4.4. Comparison of prediction results of different depth models

Several other models are tested under the same experimental conditions to provide a more accurate evaluation of the designed SSA model's water quality prediction effectiveness. These models include the traditional LSTM model, Radial Basis Function (RBF) model, differential ARIMA model, and Seq2Seq model. The comparison of prediction errors for DO is drawn in Fig. 13.

Fig. 13 illustrates the comparison of prediction errors for DO among different models. The results indicate that the designed SSA model outperforms other models regarding prediction accuracy, fitting degree, and overall prediction effectiveness.

Compared to the LSTM model, the designed SSA model shows a reduction in RMSE, MAE, and R^2 values by 39.8 %, 35.0 %, and 24.9 %, respectively. When compared to the RBF model, the SSA model exhibits a decrease in RMSE, MAE, and R^2 values by 40.5 %, 38.0 %, and 25.9 %, respectively. In comparison to the ARIMA model, the SSA model demonstrates a decrease in RMSE and MAE values by 43.5 % and 38.2 %, respectively, while R^2 increases by 69.0 %. Additionally, when compared to the Seq2Seq model, the SSA model shows a decrease in MAE and RMSE values by 31.4 % and 25.1 %, respectively, and an increase in R^2 by 14.7 %. These comparative results demonstrate the SSA model's remarkable advantages in the treatment of mine water quality time series prediction.

Despite the LSTM model having a slight advantage in capturing changes in water quality trends, the SSA model is more stable and reliable in the comprehensive performance evaluation. Although the LSTM model can capture the trend of changes in time series well, it is slightly inferior to the SSA model in terms of prediction accuracy and fit. This indicates that the model's performance under different indicators should be considered comprehensively in selecting a suitable model for a specific forecasting task. The advantages of the SSA model, which uses attention mechanisms to capture long-term dependencies, are further discussed, especially for complex sequence prediction problems. This ability enables the SSA model to more accurately predict the variation trend of water quality in diverse time scales when processing mine water quality data. Hence, it can provide vital support for water quality management and environmental protection decision-making.

The comparative outcomes of mine water quality time series prediction employing diverse DL models are illustrated in Fig. 14. Various DL models showcase disparities in their predictions of mine water quality time series. In this experimental context, the SSA model consistently presents relatively favorable comprehensive performance across key metrics, including RMSE, MAE, and coefficient of determination. This steadfast performance underscores its superior prowess in mine water quality prediction. Concurrently, the LSTM model gains a marginal edge in terms of the coefficient of determination, effectively revealing its proficiency in capturing shifts in water quality trends.

These findings consistently indicate that the designed SSA model outperforms the other tested models, offering higher accuracy, a better fitting degree, and an improved overall prediction effect. These results validate the designed SSA model's effectiveness in water quality prediction.

To further explore the performance of different DL models in the mine water quality prediction field, a series of experiments are

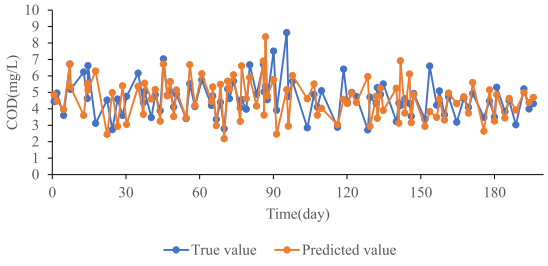


Fig. 12. Prediction results of COD.

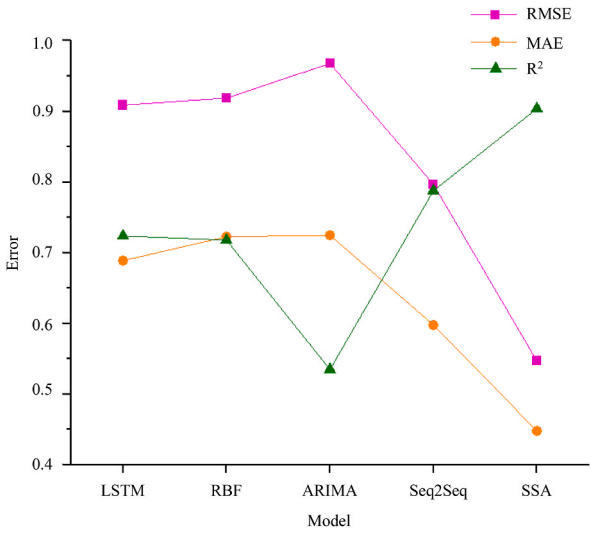


Fig. 13. Comparison of prediction errors of DO.

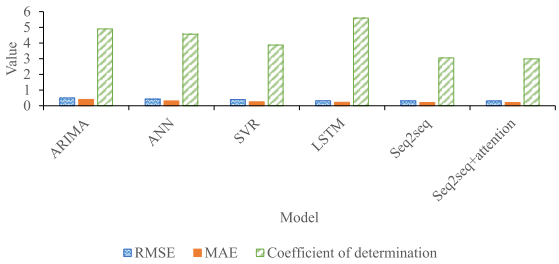


Fig. 14. Time series prediction results of mine water quality under different DL models.

conducted to compare various prediction techniques, involving classical machine learning methods and the latest DL model. These models include the SSA, the Extreme Learning Machine (ELM), the Back Propagation Neural Network (BPNN), the Temporal Convolutional Network (TCN), GRU, LSTM, and the ARIMA model. Through these experiments, the aim is to identify which models can predict water quality parameters more accurately, and then afford a scientific basis for mine water quality management.

The performance of these models is assessed against six key indicators: RMSE, MAE, R^2, Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), and Explained Variance (EV). The performance of each model examined here on the above evaluation indicators is summarized in Table 3. By comparing these results, the advantages and limitations of each model in the mine water quality prediction task can be better understood.

Table 3 compares the performance of various models in the mine water quality prediction, and their prediction ability is comprehensively evaluated through six key indicators. RMSE is an important indicator to measure the deviation between the predicted and actual values. In this research, the SSA model exhibits the lowest RMSE value (0.025), indicating that it is superior to other models regarding overall prediction accuracy. In contrast, the XGBoost and ELM models have higher RMSE values (0.030 and 0.035, respectively), showing that their prediction errors are slightly larger. MAE measures the predicted and actual values' mean absolute

Table 3
Comparison of detection performance of different models.

Model	RMSE	MAE	R^2	MAPE	MASE	EV
SSA	0.025	0.015	0.98	5.0 %	0.75	0.95
XGBoost	0.030	0.020	0.96	6.2 %	0.82	0.92
ELM	0.035	0.025	0.95	7.5 %	0.88	0.90
BP	0.040	0.030	0.93	8.3 %	0.90	0.88
GRU	0.028	0.018	0.97	5.5 %	0.78	0.94
TCN	0.026	0.017	0.97	5.2 %	0.76	0.95
LATM	0.027	0.016	0.98	5.3 %	0.77	0.96
ARIMA	0.038	0.028	0.92	9.1 %	0.92	0.85

deviation. The SSA model again shows the lowest MAE value (0.015), indicating an advantage in prediction accuracy. In contrast, the BP model's MAE value is higher (0.040), which shows that the mean absolute deviation of prediction is larger. R^2 is used to evaluate the model's ability to explain data variability, i.e. the degree of fit. The SSA model performs well in this indicator, showing an R^2 value of up to 0.98, illustrating that it can capture the variation characteristics of mine water quality data well. In contrast, the BP model exhibits a low R^2 value (0.93), demonstrating its limitations in data fitting.

MAPE measures the percentage size of the prediction error. The SSA and TCN models exhibit low MAPE values. It indicates that they excel in percentage prediction error, which is critical to ensure the practicality and accuracy of predictions. MASE measures prediction accuracy relative to the simple baseline model. The SSA and GRU models display lower MASE values, underscoring that they have better prediction accuracy compared to other models. EV is employed to assess a model's ability to interpret data variance. The SSA, LSTM, and TCN models exhibit higher EV values, which shows that they can explain the change characteristics of mine water quality data more effectively.

To sum up, the SSA model performs well on several vital performance indicators, demonstrating low prediction errors and high fit in RMSE, MAE, and R^2. In contrast, other models such as XGBoost, ELM, and BP, while doing well on some metrics, are slightly less impressive in overall performance comparisons. For the field of mine water quality prediction, it is critical to select a suitable model, and its prediction accuracy, interpretability, and applicability in practical applications should be comprehensively considered to ensure the prediction results' reliability and practicality.

Additionally, it is noted that introducing the SSA model can better capture important features in time series, which has a notable effect on improving prediction accuracy and reducing prediction errors. This finding highlights the importance of considering the model's time-dependent capture capabilities and feature concerns when conducting complex time series analyses such as water quality prediction. In short, the SSA model performs best in this research, showing strong adaptability to the task of mine water quality prediction. Future research could further explore how to optimize the attention mechanism and other parameters of high-performance models to achieve more accurate water quality predictions.

To further compare the robustness of different models, industry experts are invited to give subjective scores on four aspects: interpretability, practicality, scenario applicability, and prediction accuracy. The results are shown in Table 4:

Table 4 shows that first, the XGBoost model performs well in interpretability and practicality, thanks to its strong predictive ability and relatively good model interpretability, which makes it easier to understand and interpret in practical applications. Especially in the prediction accuracy score, XGBoost gets the highest score, showing its superiority in processing complex time series data. Second, the TCN model performs best in terms of scenario applicability. Due to its unique convolutional structure and multi-resolution imaging method, TCN can effectively capture complicated patterns and long-term dependencies in time series data. This makes the TCN model exhibit good adaptability in different application scenarios, especially in an environment that needs to deal with complex changes and rapid response.

Moreover, the SSA model's scores remain stable in all aspects, especially in practicality and scenario applicability scores. By introducing the attention mechanism, the SSA model can effectively capture the long-distance dependencies in time series, which is pivotal for complex sequence prediction tasks. Therefore, although the SSA model may be slightly inferior to XGBoost and TCN regarding interpretability and prediction accuracy, its stability and reliability in practical applications make it a strong choice. In conclusion, each model shows its advantages and characteristics under various evaluation indicators. When selecting a model suitable for a specific application, its performance in interpretability, practicality, scenario applicability, and prediction accuracy should be considered comprehensively.

As more traditional neural network models, ELM and BP may not be as interpretable as tree-based models such as XGBoost. Besides, they may not be as good as specially designed time series models such as TCN and SSA models in complex water quality prediction scenarios, so they have lower scores. GRU, as a variant of LSTM, also performs well with time series data but may be slightly inferior to SSA and TCN models in particular scenario applicability, as its model structure determines its potential limitations in dealing with very long sequences.

To sum up, each model has its advantages and disadvantages in different subjective evaluation indicators. Choosing the most suitable model not only needs to consider the prediction performance but also needs to make a comprehensive judgment according to the needs of actual application scenarios, the requirements of the model's interpretability, and the convenience of actual operation.

4.5. Discussion

In the realm of water quality prediction, this research employs a DL neural network model and introduces an attention mechanism to forecast multi-parameter water quality changes in mining contexts. The model showcases robust performance in predicting diverse water quality parameters through meticulous experimental validation. The method used in this research is compared with other similar studies to demonstrate its superiority and innovation. Gai et al. (2023) employed an optimized logistic regression algorithm for agricultural water quality prediction, achieving an average enhancement of prediction accuracy by 1.11 percentage points [38]. Although progress has been made in improving prediction accuracy, its performance in the face of multi-parameter and complex time series dependencies is relatively limited. In contrast, the proposed model can capture long-term dependencies in time series more effectively by introducing an attention mechanism, thus performing well in multi-parameter prediction.

Lv et al. (2023) harnessed water quality variables to establish an Attention-based LSTM model, predicting the water quality of the Guangzhou River section of the Pearl River with an R^2 of 0.6 [39]. Similarly, the proposed method also integrates the attention mechanism, but extends the multi-parameter prediction, demonstrating its superiority in capturing complex water quality changes. This suggests that attention mechanisms have universal applicability in enhancing models' attention to and understanding of

Table 4
Robustness comparison of diverse models.

Model	Interpretability score	Practicality score	Scenario applicability score	Prediction accuracy score
SSA	3	4	4	4
XGBoost	4	5	4	5
ELM	2	3	3	3
BP	3	3	3	3
GRU	3	4	4	4
TCN	4	4	5	4

important information in time series. Moreover, Yurtsever et al. (2023) fused SVR and XGBoost methods to create a hybrid approach, with the hybrid model achieving superior performance [40]. Nonetheless, the proposed model yields improved results in both multi-step and multi-parameter prediction, particularly with the integration of the attention mechanism, rendering it more versatile. More & Wolkersdorfer (2023) [41] proposed a solution that could help treatment plants work more efficiently and achieve their mine water management goals. It was demonstrated that AI techniques could optimize and predict mine water treatment plant parameters. However, a robust statistical analysis of the data must be performed before attempting to construct a prediction model.

In summation, this research is groundbreaking and pragmatic in the water quality prediction domain. Through comparative analysis with other research outcomes, the DL model and its attention mechanism show excellent performance in multi-parameter and multi-step water quality prediction tasks. These comparisons strengthen the model's credibility and highlight its practical application potential in water quality management and environmental protection. The proposed model stands out in multi-parameter and multi-step prediction, offering robust support for practical applications. This research not only focuses on the model's predictive performance but also discusses the model's practical application ability, especially its applicability in complex environments. Through careful comparison and analysis of different models' performance in water quality prediction, it can be found that while traditional models such as ARIMA and SVR are still effective in some situations, they face challenges in handling multi-parameter and multi-step prediction tasks, especially in scenarios where complex time series dependencies need to be captured. The attention-mechanically-enhanced DL model introduced here shows significant advantages, especially in long-term forecasting and multi-parameter forecasting, which underscores the potential of DL technology in environmental monitoring and management.

Furthermore, considering the availability of data in practical applications and the variability of the environment, the model's robustness and generalization ability are also discussed. By testing the model on the data of different seasons and regions, it is proved that the model has good generalization ability and can adapt to the water quality prediction under various environmental conditions. This finding provides vital guidance for the actual deployment of the model, showing that with proper training and adjustment, the DL model can be effectively applied to a wide range of water quality management scenarios.

5. Conclusions

In this research, a water quality prediction model based on the SSA method is developed to predict water quality in mining areas. The experimental results verify the SSA model's effectiveness in water quality prediction. The main conclusions are as follows. 1) The optimal input step of the SSA model is 40, demonstrating the lowest RMSE and MAE values and the highest R^2 values, indicating the SSA model's stability and accuracy in water quality prediction. 2) The pH value, DO, NH_3-N, and COD predicted by the SSA model are very close to the actual results, which proves its effectiveness in accurately predicting these water quality indicators, and is suitable for water quality prediction in mining areas. 3) Compared with traditional models such as ANN, LSTM, and Seq2Seq, the SSA model exhibits lower RMSE and MAE values and higher R^2 values, showing higher fit and superior prediction performance. However, a limitation of this research is the limited amount of data, which only validated the predictive performance of four physical and chemical indicators. Future studies may consider incorporating a wider range of factors, constructing more complex models to improve the accuracy of water quality predictions, and using larger and diverse data sets to further enhance the prediction model and validate its effectiveness in predicting more water quality indicators.

Data availability statement

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

CRediT authorship contribution statement

Xiaolong Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Data curation, Conceptualization. **Yang Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Formal analysis, Data curation, Conceptualization.

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.

References

[1] Y. Chen, L. Song, Y. Liu, et al., A review of the artificial neural network models for water quality prediction, Appl. Sci. 10 (17) (2020) 5776.
[2] C. Shorten, T.M. Khoshgoftaar, B. Furht, Deep learning applications for COVID-19, Journal of big Data 8 (1) (2021) 1–54.
[3] B. Lim, S. Zohren, Time-series forecasting with deep learning: a survey, Philosophical Transactions of the Royal Society A 379 (2194) (2021) 20200209.
[4] Z. Hao, W. Li, J. Wu, S. Zhang, S. Hu, A novel deep learning model for mining nonlinear dynamics in lake surface water temperature prediction, Rem. Sens. 15 (4) (2023) 900.
[5] A. Gasparin, S. Lukovic, C. Alippi, Deep learning for time series forecasting: the electric load case, CAAI Transactions on Intelligence Technology 7 (1) (2022) 1–25.
[6] S. Shweikani Debow, K. Aljoumaa, Predicting and forecasting water quality using deep learning, International Journal of Sustainable Agricultural Management and Informatics 9 (2) (2023) 114–135.
[7] S. Barra, S.M. Carta, A. Corriga, et al., Deep learning and time series-to-image encoding for financial forecasting, IEEE/CAA Journal of Automatica Sinica 7 (3) (2020) 683–692.
[8] Q. Song, Z. Wang, T. Wu, Risk analysis and assessment of water resource carrying capacity based on weighted gray model with improved entropy weighting method in the central plains region of China, Ecol. Indicat. 160 (2024) 111907.
[9] J. Dong, Z. Wang, J. Wu, J. Huang, C. Zhang, A water quality prediction model based on signal decomposition and ensemble deep learning techniques, Water Sci. Technol. 88 (10) (2023) 2611–2632.
[10] Z. Wang, Q. Wang, T. Wu, A novel hybrid model for water quality prediction based on VMD and IGOA optimized for LSTM, Front. Environ. Sci. Eng. 17 (7) (2023) 88.
[11] X. Fu, Q. Zheng, G. Jiang, K. Roy, L. Huang, C. Liu, K. Li, H. Chen, X. Song, J. Chen, Water quality prediction of copper-molybdenum mining-beneficiation wastewater based on the PSO-SVR model, Front. Environ. Sci. Eng. 17 (8) (2023) 98.
[12] E. Dritsas, M. Trigka, Efficient data-driven machine learning models for water quality prediction, Computation 11 (2) (2023) 16.
[13] B. Aslam, A. Maqsoom, A.H. Cheema, F. Ullah, A. Alharbi, M. Imran, Water quality management using hybrid machine learning and data mining algorithms: an indexing approach, IEEE Access 10 (2022) 119692–119705.
[14] W. Liu, T. Liu, Z. Liu, H. Luo, H. Pei, A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction, Environ. Res. 224 (2023) 115560.
[15] S. Talukdar, S. Ahmed, M.W. Naikoo, A. Rahman, S. Mallik, S. Ningthoujam, G.V. Ramana, Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms, J. Clean. Prod. 406 (2023) 136885.
[16] F. Xiao, Y. Cheng, P. Zhou, et al., Fabrication of novel carboxyl and amidoxime groups modified luffa fiber for highly efficient removal of uranium (VI) from uranium mine water, J. Environ. Chem. Eng. 9 (4) (2021) 105681.
[17] X. Duan, F. Ma, H. Gu, et al., Identification of mine water sources based on the spatial and chemical characteristics of bedrock brines: a case study of the xinli gold mine, Mine Water Environ. 41 (1) (2022) 126–142.
[18] M. Najafzadeh, F. Homaei, S. Mohamadi, Reliability evaluation of groundwater quality index using data-driven models, Environ. Sci. Pollut. Control Ser. 29 (6) (2022) 8174–8190.
[19] [Y. Chen, Y. Zhang, J. He, et al., Assessment of groundwater quality and pollution in the songnen plain of jilin Province, northeast China, Water 13 (17) (2021) 2414.
[20] W.C. Leong, A. Bahadori, J. Zhang, et al., Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), Int. J. River Basin Manag. 19 (2) (2021) 149–156.
[21] B. Singh, P. Sihag, V.P. Singh, et al., Soft computing technique-based prediction of water quality index, Water Supply 21 (8) (2021) 4015–4029.
[22] P. Lara-Benítez, M. Carranza-García, J.C. Riquelme, An experimental review on deep learning architectures for time series forecasting, Int. J. Neural Syst. 31 (3) (2021) 2130001.
[23] P. Nath, P. Saha, A.I. Middya, et al., Long-term time-series pollution forecast using statistical and deep learning methods, Neural Comput. Appl. 33 (19) (2021) 12551–12570.
[24] M.K. Abdel-Fattah, A. Mokhtar, A.I. Abdo, Application of neural network and time series modeling to study the suitability of drain water quality for irrigation: a case study from Egypt, Environ. Sci. Pollut. Control Ser. 28 (1) (2021) 898–914.
[25] X. Yang, J. Guan, L. Ding, et al., Research and applications of artificial neural network in pavement engineering: a state-of-the-art review, J. Traffic Transport. Eng. 8 (6) (2021) 1000–1021.
[26] J. Hagenauer, M. Helbich, A geographically weighted artificial neural network, Int. J. Geogr. Inf. Sci. 36 (2) (2022) 215–235.
[27] M. Gauch, F. Kratzert, D. Klotz, et al., Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci. 25 (4) (2021) 2045–2062.
[28] J. Qiu, B. Wang, C. Zhou, Forecasting stock prices with long-short term memory neural network based on attention mechanism, PLoS One 15 (1) (2020) e0227222.
[29] Z. Xiang, J. Yan, I. Demir, A rainfall-runoff model with LSTM-based sequence-to-sequence learning, Water resources research 56 (1) (2020) e2019WR025326.
[30] S. Ghimire, R.C. Deo, H. Wang H, et al., Stacked LSTM sequence-to-sequence autoencoder with feature selection for daily solar radiation prediction: a review and new modeling results, Energies 15 (3) (2022) 1061.
[31] H. Han, C. Choi, J. Jung, et al., Deep learning with long short term memory based Sequence-to-Sequence model for Rainfall-Runoff simulation, Water 13 (4) (2021) 437.
[32] S. Heo, H. Kim, Toward load identification based on the Hilbert transform and sequence to sequence long short-term memory, IEEE Trans. Smart Grid 12 (4) (2021) 3252–3264.
[33] R. Li, S. Zheng, C. Duan, et al., Classification of hyperspectral image based on double-branch dual-attention mechanism network, Rem. Sens. 12 (3) (2020) 582.
[34] C. Hu, Y. Wang, An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images, IEEE Trans. Ind. Electron. 67 (12) (2020) 10922–10930.
[35] H. Li, H. Duan, Y. Zheng, et al., A CTR prediction model based on user interest via attention mechanism, Appl. Intell. 50 (4) (2020) 1192–1203.
[36] A. Surov, H.J. Meyer, A. Wienke, Correlations between apparent diffusion coefficient and gleason score in prostate cancer: a systematic review, European Urology Oncology 3 (4) (2020) 489–497.
[37] D.L. Guenaga, O.E. Marcillo, A.A. Velasco, et al., The silencing of US campuses following the COVID-19 response: evaluating root mean square seismic amplitudes using power spectral density data, Seismol Res. Lett. 92 (2A) (2021) 941–950.
[38] R. Gai, H. Zhang, Prediction model of agricultural water quality based on optimized logistic regression algorithm, EURASIP Journal on Advances in Signal Processing 2023 (1) (2023) 21.
[39] M. Lv, X. Niu, D. Zhang, H. Ding, Z. Lin, S. Zhou, Y. Zhu, A data-driven framework for spatiotemporal analysis and prediction of river water quality: a case study in Pearl River, China, Water 15 (2) (2023) 257.
[40] M. Yurtsever, E. Murat, Potable water quality prediction using artificial intelligence and machine learning algorithms for better sustainability, Ege Academic Review 23 (2) (2023) 265–278.
[41] K.S. More, C. Wolkersdorfer, Application of machine learning algorithms for nonlinear system forecasting through analytics—a case study with mining influenced water data, Water Resour. Ind. 29 (2023) 100209.