



OPEN ACCESS

EDITED BY

Arindam Malakar,
University of Nebraska-Lincoln, United States

REVIEWED BY

Alakananda Mitra,
University of Nebraska-Lincoln, United States
Muhammad Hamza Zafar,
University of Agder, Norway

*CORRESPONDENCE

Yajie Liang,
✉ liangyajie202404@163.com

RECEIVED 29 May 2024

ACCEPTED 15 July 2024

PUBLISHED 08 August 2024

CITATION

Liang Y, Zhao J, Zhang Y, Li J, Ding J, Jing C, Ji J and Wu D (2024), Developing an SSA-optimized attention-ConvGRU model for predicting and assessing soil contaminant distribution. *Front. Environ. Sci.* 12:1440296. doi: 10.3389/fenvs.2024.1440296

COPYRIGHT

© 2024 Liang, Zhao, Zhang, Li, Ding, Jing, Ji and Wu. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Developing an SSA-optimized attention-ConvGRU model for predicting and assessing soil contaminant distribution

Yajie Liang^{1,2*}, Jieyu Zhao^{1,2}, Yiting Zhang^{1,2}, Jisheng Li³, Jieran Ding^{1,2}, Changyong Jing^{1,2}, Jiukun Ji¹ and Dongtan Wu¹

¹Department of Environmental Engineering, Hebei University of Environmental Engineering, Qinhuangdao, China, ²Hebei University of Environmental Engineering, Hebei Key Laboratory of Agroecological Safety, Qinhuangdao, China, ³Central Control Technology (Xi'an) Co., Ltd., Xi'an, China

Introduction: Soil pollution, which includes a variety of contaminants such as heavy metals and organic compounds, poses significant environmental and health risks, making effective prediction and assessment techniques essential. Current predictive models often struggle with the complexity and diversity of soil contaminant behaviors, leading to limitations in their accuracy and applicability.

Methods: Recognizing the importance of capturing the temporal dynamics influenced by seasonal variations and agricultural practices, our study introduces an SSA-optimized Attention-ConvGRU model. This model integrates convolutional neural networks, gated recurrent units, and attention mechanisms, enhanced through optimization with the Sparrow Search Algorithm to improve predictive performance.

Results: Experimental results confirm that our model significantly outperforms traditional methods, demonstrating over 30% improvement in prediction accuracy across multiple datasets.

Discussion: This research underscores the potential of advanced machine learning techniques to revolutionize the assessment of soil pollution, providing substantial benefits for environmental management and public health protection.

KEYWORDS

soil pollution, deep learning, sparrow search algorithm, ConvGRU, attention mechanisms, attention-ConvGRU

1 Introduction

Soil is an essential component of the Earth's system, providing support for human production and life through its production, regulation, cycling, and habitat functions. It plays a critical role in environmental regulation by purifying pollutants to protect the environment [Gautam et al. \(2023\)](#); [Raimi et al. \(2022\)](#). However, rapid industrialization and urbanization have led to severe soil pollution, threatening ecosystems and human health. Soil pollutants include various chemicals such as heavy metals, organic pollutants, and pesticide residues, originating from industrial production, agricultural activities, urbanization processes, and other human activities [Wu et al. \(2022\)](#); [Ren and Wang \(2024\)](#). Soil pollution has a wide range and a broad impact, and its consequences may include declining soil quality, inhibited

plant growth, groundwater pollution, and human health issues. Therefore, effectively predicting and assessing the distribution and diffusion of soil pollutants has become an important task for environmental protection and human health. Despite extensive research in the field of soil pollution, the diversity in the nature and behavior of soil pollutants, along with the complexity of the soil environment, poses significant challenges in predicting and assessing soil pollution Wang et al. (2022, 2024). Different pollutants interact and diffuse in soil in various ways, and factors such as soil heterogeneity, seasonal variations, and topographic features further complicate the development of effective prediction models Liu et al. (2023). With the development of deep learning technology, researchers have begun to utilize deep learning methods to study soil pollution issues. Compared to traditional statistical methods and machine learning methods, deep learning has stronger modeling capabilities and better generalization abilities, allowing it to learn complex nonlinear relationships from large amounts of data Gao et al. (2022); Janga et al. (2023). Therefore, deep learning has shown tremendous potential in the prediction and assessment of soil pollutants. Researchers have proposed numerous deep learning-based models to predict the distribution, diffusion, and risk assessment of soil pollutants. These models not only enhance the accuracy of predictions but also speed up the prediction process, providing crucial support for the management and remediation of soil pollution issues Tao et al. (2022); Jiang et al. (2022).

Researchers have proposed various deep learning-based models for predicting analysis and risk assessment of soil pollutants. One of these is based on Convolutional Neural Networks (CNN), which extracts features from spatiotemporal data of soil pollutants through convolutional operations for prediction Mei et al. (2022). While CNNs have certain advantages in feature extraction, their structural characteristics may overlook temporal information in the data, thereby limiting the accuracy of predictions to some extent. Another study employs Long Short-Term Memory networks (LSTM) as the foundational model Zheng et al. (2022). LSTMs play a crucial role in enhancing predictions of the spatiotemporal distribution of soil pollutants by effectively capturing long-term dependencies in time series data. However, despite partially addressing temporal dependencies, LSTM's modeling capability for long-term dependencies may still be insufficient, resulting in certain limitations in predictions. Another approach involves using Variational Autoencoders (VAE) for soil pollutant prediction Shin et al. (2023). VAEs learn the latent representation of data and generate new samples with similar distributions. While this model has advantages in capturing data distributions, its performance in handling soil pollutants may be affected by data sparsity and noise, leading to decreased prediction accuracy. The last study combines attention mechanisms and Gated Recurrent Units (GRU) Mirzavand Borujeni et al. (2023). This model dynamically weights input data using attention mechanisms and captures long-term dependencies in sequence data using GRU. While this model improves prediction performance to some extent, its modeling of complex relationships between different types of pollutants still requires improvement.

In summary, these studies have made progress in predicting soil pollutants, but there are still some shortcomings. Models may overlook crucial temporal information in spatiotemporal data, and there is room for improvement in modeling long-term dependencies and the complex relationships between different types of pollutants Cui et al. (2023). Therefore, further research and improvement of deep learning models

are needed to accurately predict the spatiotemporal distribution of soil pollutants and provide more effective support for environmental protection and soil remediation.

Although these studies have made some progress in predicting soil pollutants, there are still some shortcomings. Current models may overlook critical temporal information when handling spatiotemporal data, have limited capability in modeling long-term dependencies, and need improvement in capturing the complex relationships between different types of pollutants. Therefore, further research and refinement of deep learning models are needed to more accurately predict the spatiotemporal distribution of soil pollutants and provide more effective support for environmental protection and soil remediation.

Based on the identified shortcomings, we propose an attention-ConvGRU model optimized with the Sparrow Search Algorithm (SSA). This model utilizes SSA to optimize parameter settings, aiming to enhance model performance and convergence speed. Additionally, the model integrates attention mechanisms, convolutional Gated Recurrent Unit (ConvGRU) to address challenges in soil pollutant prediction. Within this model, attention mechanisms dynamically weight input data, allowing the model to focus on information relevant to the prediction target, thus improving prediction accuracy. CNNs are employed to extract spatiotemporal features from input data, aiding in capturing spatial distributions and temporal variations of soil pollutants. GRUs handle time-series data, proficiently capturing long-term dependencies among data points and enhancing the model's ability to model temporal dynamics.

The introduction of our model holds significant importance. By leveraging the SSA algorithm to optimize model parameters, we are able to accelerate the model's convergence process, thereby enhancing training efficiency. Moreover, the integration of attention mechanisms, convolutional neural networks, and gated recurrent units enables the model to fully exploit the spatiotemporal information of soil pollutant data, leading to improved prediction accuracy and stability. Consequently, our model shows promising prospects for applications in soil pollutant prediction and risk assessment, offering more effective support for environmental protection and soil remediation efforts.

The SSA-optimized attention-ConvGRU model contributes significantly to three aspects in the field of soil pollutant prediction and risk assessment.

- We introduced a method to optimize model parameters using the SSA algorithm, which effectively enhances the model's performance and training speed. The incorporation of the SSA algorithm accelerates the model's convergence process, enabling it to learn effective feature representations from data more rapidly, thus improving prediction accuracy.
- We integrated three model structures - attention mechanisms, convolutional neural networks, and gated recurrent units - organically, fully leveraging their respective advantages. The attention mechanism dynamically focuses on information relevant to the prediction target, the convolutional neural network extracts spatiotemporal features, and the gated recurrent unit effectively captures long-term dependencies among data points. The integration of these three elements empowers the model to make more precise predictions regarding the spatiotemporal distribution of soil pollutants.

- The proposed model achieved significant predictive performance in experiments, outperforming traditional methods and other deep learning models. This indicates that the method we proposed has high practical value in soil pollutant prediction and risk assessment, providing important technical support for research and practice in related fields.

2 Related work

2.1 Traditional methods for predicting soil pollutants

In terms of traditional methods for predicting soil pollutants, previous scholars have employed various approaches for research and prediction. These methods often rely on empirical models [Zeng et al. \(2022\)](#), statistical techniques [Baragaño et al. \(2022\)](#), and Geographic Information System (GIS) technology [Khan et al. \(2022\)](#) to forecast the distribution and diffusion of soil pollutants. Empirical models are a common approach, which are based on historical data and expert knowledge to predict future soil pollution scenarios by summarizing and generalizing past experiences. However, empirical models often lack adaptability and precision when dealing with complex, non-linear interactions among multiple soil pollutants and environmental factors. Statistical methods are also widely used, including regression analysis [Bijitha and Nath \(2022\)](#); [Zhao et al. \(2023\)](#), variance analysis, cluster analysis, etc., to establish mathematical models describing the distribution patterns and influencing factors of soil pollutants through statistical analysis of existing data. Despite their usefulness, statistical methods can be limited by the quality and quantity of available data and may not fully capture the dynamic and heterogeneous nature of soil pollution. Furthermore, Geographic Information System (GIS) technology plays a crucial role in soil pollutant prediction, integrating spatial data, topographic information, soil properties, and other multi-source data to build spatial databases and conduct spatial analysis, thereby revealing distribution patterns and spatial correlations of soil pollutants [Xu and Zhang \(2023\)](#); [Liu et al. \(2022\)](#). GIS-based models, while powerful in spatial analysis, often require extensive and high-quality spatial data, which can be challenging to obtain and maintain. Additionally, geostatistical methods combine principles of geology and statistics to model and predict the spatial distribution of soil pollutants [Yao and Liu \(2024\)](#). These methods utilize statistical analysis of geological data to explore the spatial variations of soil pollutants and make spatial predictions [Chelabi et al. \(2022\)](#). However, geostatistical methods may struggle with the spatial autocorrelation and non-stationarity often present in soil pollution data.

In summary, traditional methods for predicting soil pollutants encompass various techniques, including empirical models, statistical methods, GIS technology, and geostatistical methods. While these methods play an important role in revealing distribution patterns of soil pollutants and forecasting future trends, they each have inherent limitations that can affect their accuracy and applicability. These limitations include issues with adaptability, data quality requirements, and the ability to capture complex interactions and dynamics in soil pollution. Therefore, there is a need for more advanced and flexible modeling approaches that can address these challenges and provide more reliable predictions for soil environmental management and protection.

2.2 Soil pollutant prediction based on convolutional neural network

In past research, soil pollutant prediction based on Convolutional Neural Networks (CNNs) has garnered widespread attention [Li et al. \(2022\)](#). CNNs, as powerful deep learning models, have achieved significant success in fields such as image processing and spatial data analysis [Pisal and Vidyarthi \(2023\)](#). When applied to soil pollutant prediction, CNNs excel at effectively extracting features from spatiotemporal soil data, enabling precise prediction of soil pollutants.

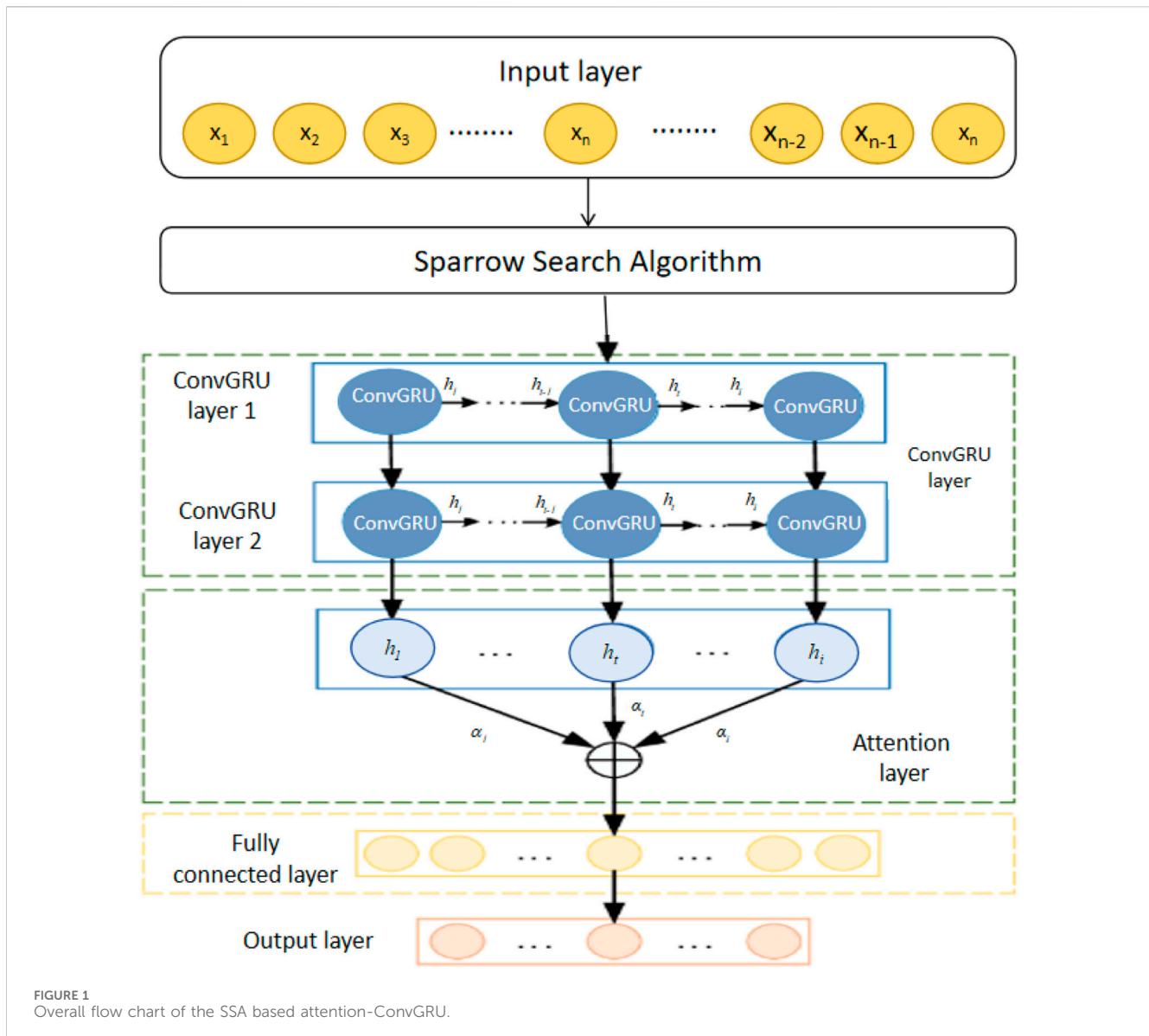
CNN models firstly leverage multiple layers of convolution and pooling operations to automatically learn features from spatiotemporal soil data [Zhang et al. \(2023\)](#). These features, including soil properties, topographical features, and environmental factors, help the model better understand the distribution patterns of soil pollutants [Siddhan and Shanthi \(2022\)](#); [Liu et al. \(2023\)](#). By extracting these features, CNNs can uncover hidden patterns and trends within large-scale soil data, thereby improving prediction accuracy and reliability. However, CNNs primarily focus on local spatial features and may struggle to capture long-range dependencies in the data. This limitation can reduce the accuracy of predictions in complex soil environments where such dependencies are significant. Secondly, CNN models possess strong spatial awareness and local connectivity, allowing them to fully consider the spatial distribution characteristics of soil data [Nadiri et al. \(2022\)](#). This enables CNNs to better capture spatial variations of soil pollutants, leading to more accurate predictions of soil pollution under different regional and terrain conditions. Nonetheless, the performance of CNNs can be highly dependent on the quality and quantity of training data, which might not always be available, potentially leading to overfitting or underfitting issues. Moreover, CNNs can be combined with other deep learning models such as Recurrent Neural Networks (RNNs) or Gated Recurrent Units (GRUs) to further enhance model performance [Gonzalez et al. \(2022\)](#); [Duan et al. \(2022\)](#). By integrating multiple deep learning models, their respective strengths can be fully utilized, enabling more precise prediction and comprehensive analysis of soil pollutants.

In conclusion, while soil pollutant prediction based on Convolutional Neural Networks holds significant potential and practical value, it is important to address the inherent limitations of CNNs. These include challenges in capturing long-range dependencies, dependency on high-quality data, and computational complexity. By leveraging information from spatiotemporal soil data and combining powerful feature extraction and spatial awareness capabilities, CNN models can provide more effective methods and tools for predicting and evaluating soil pollutants, thus offering crucial support for environmental protection and soil remediation efforts. Future research should focus on overcoming these limitations to fully realize the potential of CNN-based models in this domain.

3 Methods

3.1 Overview of our network

This study proposes an attention-ConvGRU model optimized with the Sparrow Search Algorithm (SSA) for predicting and



assessing the risk of soil pollutants. This model integrates attention mechanisms, Convolutional Gated Recurrent Units (ConvGRU), and the Sparrow Search Algorithm (SSA). The attention mechanism enables the model to focus more on features relevant to soil pollution prediction, thereby enhancing accuracy and generalization. ConvGRU captures the spatiotemporal characteristics of soil pollution data, aiding the model in effectively analyzing the spatial distribution and temporal changes of soil pollution. The Sparrow Search Algorithm is employed for optimizing model parameters, improving training efficiency and prediction accuracy. In the process of network construction, soil data is first processed through the attention mechanism to extract important features. These characteristics are subsequently input into the ConvGRU unit to capture temporal connections and spatial interdependencies within the data. Subsequently, SSA is employed to optimize the weights and biases of the entire network, ensuring that the model achieves optimal prediction performance after multiple iterations of training. Through this

approach, the model not only enhances its ability to identify existing pollution levels but also improves its capability to predict future pollution trends. The overall flowchart is illustrated in [Figure 1](#).

The application of this model is significant for soil pollutant detection as it accurately predicts the concentration changes of various pollutants in soil. This aids environmental scientists and policymakers in more effectively assessing soil pollution risks and formulating corresponding prevention and control measures. By deeply analyzing the spatial distribution and temporal trends of soil pollution, this model contributes to more accurately guiding the implementation of soil remediation and pollution prevention strategies.

3.2 Sparrow Search Algorithm

The Sparrow Search Algorithm (SSA) is a heuristic optimization algorithm inspired by the foraging behavior of sparrows. In this

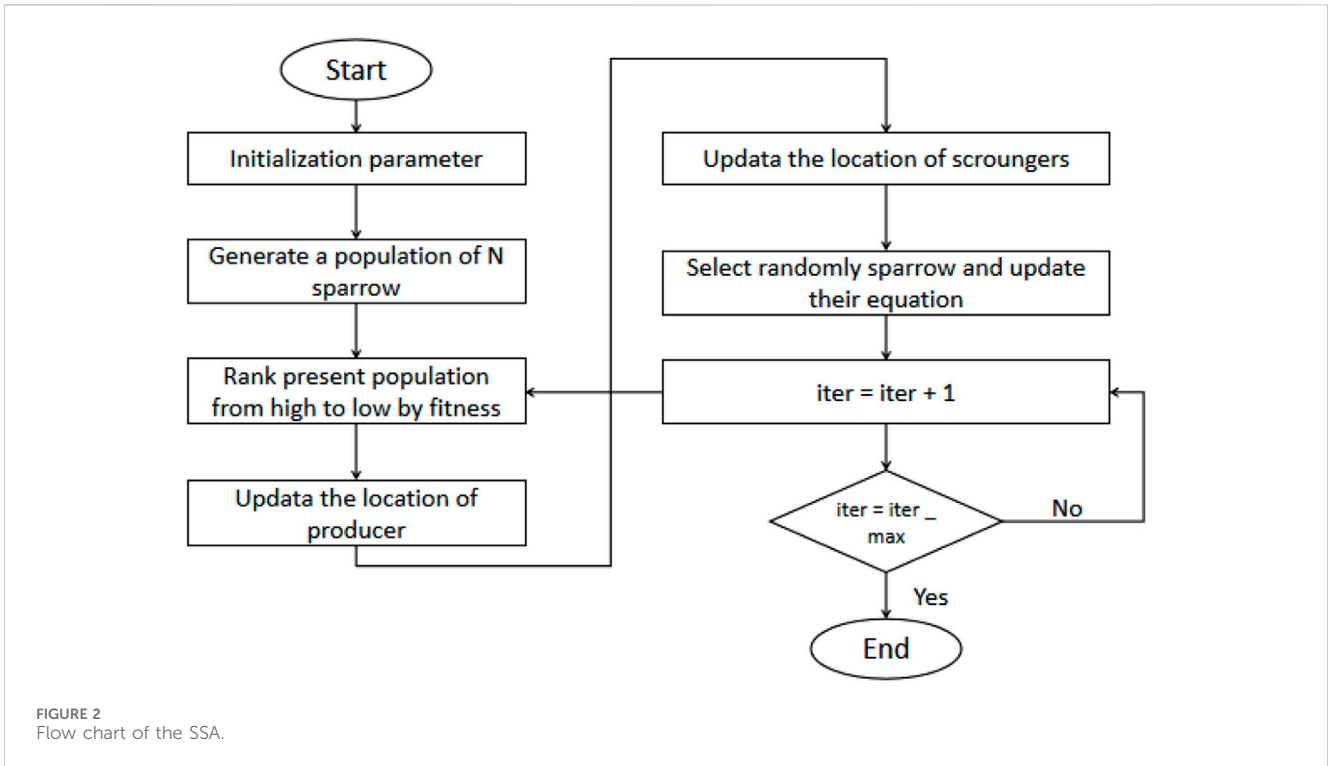


TABLE 1 Core mathematical formulas related to SSA.

Formula type	Mathematical formula	Description
Objective Function	$f(x) = \sum_{i=1}^n x_i^2(1)$	$f(x)$ is the objective function to be minimized. Here, $x = (x_1, x_2, \dots, x_n)$ represents the solution vector
Exploration Step	$\Delta x_i = \beta \cdot (x_{max} - x_{min}) \cdot (r - 0.5)(2)$	Δx_i represents the exploration step size, β is a parameter controlling the step size, x_{max} and x_{min} are the bounds of the solution space, r is a random number in the range $[0, 1]$
Update Rule for Position	$x_i(t + 1) = x_i(t) + r \cdot (x_{best}(t) - x_i(t)) + \alpha \cdot \Delta x_i(3)$	$x_i(t)$ is the position of the i -th sparrow at time t , $x_{best}(t)$ is the position of the best sparrow at time t , r is a random number in the range $[0, 1]$, α is a step size parameter, Δx_i is a random vector representing the exploration step
Pursuing Step	$x_i(t + 1) = x_i(t) + \delta \cdot (x_{leader}(t) - x_i(t))(4)$	$x_i(t)$ is the position of the i -th sparrow at time t , $x_{leader}(t)$ is the position of the leading sparrow at time t , δ is a parameter controlling the influence of the leader
Scouting Behavior	$x_i(t + 1) = x_i(t) + \lambda \cdot (r - 0.5)(5)$	$x_i(t)$ is the position of the i -th sparrow at time t , λ is a parameter controlling the scouting behavior, r is a random number in the range $[0, 1]$
Calculation of Fitness	$Fit_i = \frac{1}{1+f(x_i)}(6)$	Fit_i is the fitness value of the i -th sparrow, $f(x_i)$ is the objective function value of the i -th sparrow
Update Rule for Leader Sparrow	$x_{best}(t + 1) = x_{best}(t) + \gamma \cdot \frac{\sum_{i=1}^N Fit_i \cdot x_i}{\sum_{i=1}^N Fit_i}(7)$	$x_{best}(t)$ is the position of the best sparrow at time t , γ is a scaling factor, N is the total number of sparrows
Stopping Criterion	Stop if $t = t_{max}(8)$	t is the current iteration, t_{max} is the maximum number of iterations

algorithm, sparrows represent candidate solutions, and the sparrows within a population collectively search the solution space of a problem. SSA achieves efficient global search and parameter optimization by simulating both individual behaviors and collective cooperation among sparrows during foraging Gharehchopogh et al. (2023). The foraging process in SSA consists of four stages: searching, pursuing, tracking, and

scouting. During the searching stage, sparrows select food sources based on their own experience and information from neighboring sparrows, adjusting their search direction according to the richness of the food Awadallah et al. (2023). In the pursuing stage, sparrows follow companions that discover more food and attempt to find additional food in the same direction. During the tracking stage, sparrows adjust their flight speed and direction to

better follow other sparrows. Finally, in the scouting stage, sparrows explore new food sources through random flights to ensure diversity and exploration within the entire population. By simulating these behaviors, SSA efficiently conducts global search and parameter optimization. In optimization problems, SSA effectively explores the solution space and finds better solutions in shorter timeframes. The algorithmic flowchart for SSA is depicted in Figure 2. The core mathematical formulas related to SSA are presented in Table 1 below.

These formulas outline the key components and operations of the SSA algorithm, including the objective function, position update, fitness calculation, leader sparrow update, and stopping criterion.

In our model, the SSA algorithm is utilized to optimize the parameters of the entire model, including the weights of CNN, GRU, and attention mechanisms, as well as the selection of hyperparameters. Through optimization with the SSA algorithm, we can quickly find the optimal parameter configuration for the model, thereby improving both training speed and performance. Moreover, the SSA algorithm aids in better exploration of the parameter space, leading to more stable and reliable models. This enhances the accuracy and robustness of our model in detecting soil pollutants. Thus, the SSA algorithm plays a crucial role in our model, providing an effective means for optimization and performance enhancement.

3.3 Attention mechanism

Attention is a pivotal mechanism designed to enhance the capacity of neural networks in processing sequential data. This mechanism allows the network to selectively focus on the most relevant segments of the input sequence, which in turn significantly improves the model's performance and accuracy. The fundamental principle of the attention mechanism involves calculating a set of weights for each position in the input sequence Guo et al. (2022). These weights reflect the importance of each part of the sequence in relation to the task at hand. By dynamically adjusting its focus based on these learned weights, the model can prioritize more critical information while disregarding less relevant data. This selective focus enables the model to better capture and understand complex patterns within the sequence, thereby enhancing its overall effectiveness in tasks such as language translation, image captioning, and time-series prediction. The attention mechanism operates by computing a weighted sum of all positions in the input sequence, where the weights are determined by the similarity between a query vector (representing the current position) and key vectors (representing all positions in the sequence). This approach allows the model to flexibly integrate contextual information from the entire sequence, leading to more informed and accurate predictions.

Query-Attention Score: The query-attention score measures the similarity between a query vector and a key vector.

$$\text{Attention}(q, k) = q \cdot k^T \quad (1)$$

where: q is the query vector, k is the key vector.

Softmax Attention Weights: The softmax attention weights determine the importance of each key based on their attention scores.

$$\text{Softmax}(s_i) = \frac{e^{s_i}}{\sum_j e^{s_j}} \quad (2)$$

where: s_i is the attention score for the i -th key.

Weighted Sum of Values: The weighted sum of values combines the values with their corresponding softmax attention weights.

$$\text{Attention_Output} = \sum_i \text{Softmax}(s_i) \cdot v_i \quad (3)$$

where: $\text{Softmax}(s_i)$ is the softmax attention weight for the i -th key, v_i is the value vector corresponding to the i -th key.

Scaled dot-product attention: Scaled dot-product attention is a specific type of attention mechanism that computes the attention scores using the dot product of query and key vectors, scaled by a factor to prevent extremely large values that can hinder the softmax function's performance. It is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimensionality of the keys. The scaling factor $\sqrt{d_k}$ ensures that the dot product values remain in a range that is conducive to effective learning. Specifically, the scaled attention score is calculated as:

$$\frac{q_i \cdot k_j^T}{\sqrt{d_k}} \quad (5)$$

where q_i is the i -th query vector, k_j is the j -th key vector, and d_k is the dimensionality of the key vectors.

Multi-Head Attention: Multi-head attention computes multiple attention heads in parallel, allowing the model to focus on different parts of the input.

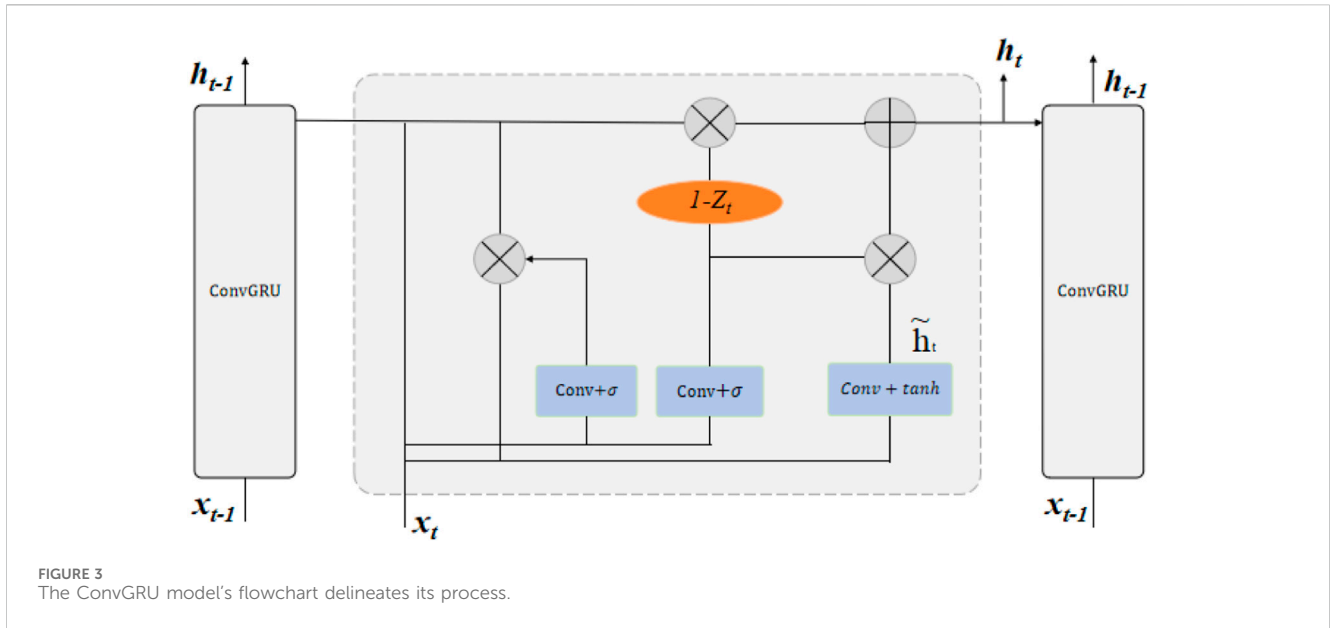
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (6)$$

where: $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, W_i^Q is the weight matrix for query in the i -th head, W_i^K is the weight matrix for key in the i -th head, W_i^V is the weight matrix for value in the i -th head, W^O is the output projection matrix.

In our model, Attention is applied to the task of predicting soil pollutants. By incorporating the Attention mechanism, our model can more flexibly identify and utilize key information within the input data, thereby boosting prediction accuracy and robustness. Particularly, Attention enables the model to prioritize regions of the input sequence that have a significant impact on the prediction of soil pollutants while effectively suppressing irrelevant information. As a result, the model can more accurately capture the distribution patterns and trends of soil pollutants, contributing to a more precise understanding of their spatial and temporal variations.

3.4 ConvGRU model

The Convolutional Gated Recurrent Unit (ConvGRU) is a model that integrates Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs). Its basic principle involves using CNN's convolutional operations to extract spatial features from sequential data and combining GRU's gated mechanism to



capture temporal features, enabling effective modeling and prediction of sequential data Zhou et al. (2024). In the ConvGRU model, convolutional layers are utilized to extract spatial features from the input data. Convolutional operations enable the model to capture local features and spatial relationships within the data efficiently. These operations are particularly effective for processing data with spatial structures, such as images or spatially organized sequences. On the other hand, GRUs are employed to manage temporal information within the sequential data. GRUs use a gated mechanism to learn and retain long-term dependencies between data points. This mechanism involves reset and update gates that control the flow of information, allowing the model to maintain relevant historical information and discard irrelevant data, thus enhancing the model's ability to predict future trends in the sequential data. Figure 3 provides a visual representation of the flowchart outlining the ConvGRU procedure.

The equations presented below outline the key components and operations of the ConvGRU model, elucidating its mechanism for processing sequential data. Each equation contributes to the understanding of how information flows through the model, culminating in the generation of output predictions.

Input Convolution: For each time step t , a convolutional operation is performed on the input x_t to extract spatial features.

$$f_t = \text{Conv}(x_t) \tag{7}$$

where f_t represents the feature map obtained from the convolutional layer.

Reset Gate: Controls the degree to which the previous hidden state h_{t-1} is reset or updated based on the current input x_t . It determines how much of the past information should be forgotten and how much of the current input should be incorporated into the current hidden state. The reset gate is computed using a sigmoid activation function σ , ensuring its value lies between 0 and 1.

$$r_t = \sigma(W_r * [h_{t-1}, x_t] + b_r) \tag{8}$$

where r_t represents the reset gate at time step t , W_r is the weight matrix of the reset gate, $*$ denotes the convolution operation, and b_r is the bias vector of the reset gate.

Update Gate: Regulates the amount of information from the previous hidden state h_{t-1} that is preserved or discarded in favor of the candidate hidden state. It governs the flow of information from the past to the present, deciding the extent to which past information should be retained and how much new information should be incorporated. Similar to the reset gate, the update gate is computed using a sigmoid activation function σ .

$$z_t = \sigma(W_z * [h_{t-1}, x_t] + b_z) \tag{9}$$

where z_t denotes the update gate at time step t , W_z is the weight matrix of the update gate, $*$ denotes the convolution operation, and b_z is the bias vector of the update gate.

Candidate Hidden State: Represents the proposed new hidden state value at time step t . It combines information from the reset gate-applied previous hidden state and the current input to generate a new candidate hidden state. The candidate hidden state is computed using the hyperbolic tangent function \tanh to ensure its values lie between -1 and 1 .

$$\tilde{h}_t = \tanh(W * [r_t * h_{t-1}, x_t] + b) \tag{10}$$

where \tilde{h}_t represents the candidate hidden state at time step t , W is the weight matrix of the candidate hidden state, $*$ denotes the convolution operation, and b is the bias vector of the candidate hidden state.

Hidden State Update: Combines information from the previous hidden state and the candidate hidden state to generate the updated hidden state at time step t . It determines how much of the past information should be retained and how much of the new information should be incorporated into the current hidden state, based on the update gate values.

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{11}$$

where h_t denotes the updated hidden state at time step t .

Layer Normalization: Layer normalization is applied to stabilize the training process and improve the model's performance. It normalizes the hidden states across the feature dimension.

$$\hat{h}_t = \frac{h_t - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (12)$$

where \hat{h}_t represents the normalized hidden state, μ and σ are the mean and standard deviation of the hidden states, and ϵ is a small constant to prevent division by zero.

Output Computation: Computes the output at time step t based on the current hidden state. It transforms the convolutional output of the hidden state into a probability distribution over the possible output classes using the softmax function.

$$y_t = \text{softmax}(V * h_t + c) \quad (13)$$

where y_t represents the output at time step t , V is the weight matrix of the output layer, $*$ denotes the convolution operation, and c is the bias vector of the output layer.

In our model, the Convolutional Gated Recurrent Unit (ConvGRU) is applied to the task of predicting soil pollutants. By incorporating the ConvGRU model, we can effectively utilize the spatiotemporal information present in soil pollutant data. Through the combination of convolutional operations and gated mechanisms, the ConvGRU model enables precise prediction and assessment of soil pollutants. Specifically, the ConvGRU model excels in capturing the spatial distribution patterns and temporal trends within soil pollutant data. This enhances the model's prediction accuracy and robustness, providing valuable technical support for soil environmental monitoring and management.

4 Experiment

4.1 Datasets

In conducting our deep learning experiments for this study, we selected four datasets:

The United States Environmental Protection Agency (EPA) dataset [Mitra et al. \(2023\)](#) provides extensive soil pollution datasets covering soil quality information nationwide. These datasets include concentrations of organic and inorganic compounds, heavy metal contents, volatile organic compounds (VOCs), and persistent organic pollutants (POPs), among others. EPA datasets typically contain detailed geographical coordinates, soil sampling information, and pollutant concentrations, which can be utilized for analyses of soil pollution distribution patterns, risk assessments, and environmental monitoring applications. The EPA dataset consists of approximately 10,000 sampling points, with each point providing concentrations for over 50 different pollutants. The dataset includes metadata such as sampling depth, soil type, and land use, which are essential for comprehensive soil quality analysis. These datasets serve as valuable references for researchers and policymakers to better understand soil pollution issues and develop corresponding management measures.

The SoilGrids dataset is a global soil property dataset [Dinamarca et al. \(2023\)](#) developed by the International Soil Reference and Information Centre (ISRIC). It provides spatially explicit

information on various soil properties, including soil organic carbon content, pH, texture, and other key attributes, at different depths and resolutions. The dataset is generated through machine learning techniques and soil observations from various sources, offering valuable insights into soil variability and distribution patterns worldwide. SoilGrids facilitates a wide range of applications, such as soil mapping, environmental modeling, agricultural management, and land use planning. It covers different soil depths (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm) with a spatial resolution of 250 m. The dataset is generated through machine learning techniques and soil observations from various sources, offering valuable insights into soil variability and distribution patterns worldwide. SoilGrids facilitates a wide range of applications, such as soil mapping, environmental modeling, agricultural management, and land use planning, contributing to improved soil management practices and sustainable land use strategies on a global scale.

The European Soil Data Centre (ESDAC) dataset [Panagos et al. \(2022\)](#) is a resource supported by the European Union (EU), aimed at providing comprehensive information about soils across Europe. This dataset contains soil sample data from various locations in Europe, covering key soil properties such as soil texture, soil type, organic carbon content, pH value, and more. These data are collected from diverse sources including ground observations, satellite remote sensing, field surveys, and soil monitoring projects in different countries and regions. The dataset includes detailed metadata for each sample point, such as geographical coordinates, sampling methods, and analysis techniques. These data are collected from diverse sources including ground observations, satellite remote sensing, field surveys, and soil monitoring projects in different countries and regions. The ESDAC dataset serves as a valuable resource of soil information for researchers, policymakers, and decision-makers, facilitating applications in soil management, environmental protection, agricultural production, and land planning.

The National Soil Inventory (NSI) dataset [Prout et al. \(2022\)](#) is a comprehensive collection of soil data compiled as part of a national-level soil survey initiative. It typically includes detailed information on soil properties, chemical composition, and contaminant concentrations from soil samples collected across various regions within a country. The NSI dataset encompasses data from thousands of sampling locations, providing extensive coverage of different soil types and land uses. It includes metadata such as sampling dates, depths, and methods, which are crucial for accurate analysis and interpretation. The NSI dataset serves as a valuable resource for soil research, environmental management, agricultural planning, and policy-making. It provides essential insights into soil quality, pollution levels, and spatial distribution patterns, supporting diverse applications in soil conservation, land use planning, and pollution mitigation efforts.

4.2 Experimental details

Step1: Data Preprocessing

- **Data Cleaning:** Involves removing or correcting any inaccuracies, inconsistencies, or missing values in the

dataset. Removing duplicate entries, handling missing data by imputation or deletion, and correcting errors in data entry. In our study, we removed entries with missing values exceeding 10% of the total dataset.

- **Data Standardization:** Standardizing the scale of the features to ensure that all variables have a similar range of values. This typically involves techniques such as mean normalization or feature scaling. We standardized the numerical features through z-score normalization to center them around a mean of 0 and scale them to have a standard deviation of 1.
- **Data Splitting:** Dividing the dataset into training, validation, and testing sets to evaluate the performance of the model. In our experiment, we split the dataset into 80% training, 10% validation, and 10% testing sets to train the model, tune hyperparameters, and evaluate its performance, respectively.
- **Feature Engineering:** It involves preprocessing the data to extract relevant features that enhance the model's predictive power. This includes selecting informative features, transforming variables to better represent relationships, and scaling features to ensure uniformity. For our study, we employed methods like feature selection, transformation, and scaling to optimize the dataset for training our soil pollutant prediction model.

Step2: Model Training

Before diving into model training, it is essential to meticulously configure network parameters, design an optimal architecture, and implement an effective training process.

- **Network Parameter Settings:** Setting the parameters of the neural network architecture is crucial for effective model training. We configured parameters such as the learning rate, batch size, and number of epochs. We set the learning rate to 0.001, the batch size to 32, and conducted training for 100 epochs to optimize the model's performance.
- **Model Architecture Design:** Designing the architecture of the neural network involves determining the number of layers, types of activation functions, and the size of hidden units. We designed a deep learning architecture with three convolutional layers followed by two GRU layers, each with a ReLU activation function. The convolutional layers had 64, 128, and 256 filters, respectively, while the GRU layers had 128 hidden units.
- **Model Training Process:** The model training process involves feeding the preprocessed data into the neural network and iteratively updating the model parameters to minimize the loss function. We utilized the Adam optimizer with a categorical cross-entropy loss function to train the model. The training process involved forward and backward propagation of data batches through the network, adjusting the weights and biases using gradient descent. We monitored the training progress by evaluating the model's performance on the validation set after each epoch, ensuring convergence and preventing overfitting.

Algorithm 1 outlines the algorithmic process for training the SSA based attention-ConvGRU network.

```

Data: EPA dataset, SoilGrids dataset, ESDAC dataset,
NSI dataset
Result: Trained SSA based attention-ConvGRU network
Initialize network parameters: learning rate, batch
size, number of epochs;
for each epoch do
  for each mini-batch in training data do
    Perform data augmentation if necessary;
    Forward pass through the network to compute
    predicted outputs;
    Compute loss function using predicted outputs and
    ground truth labels;
    Backpropagate gradients through the network;
    Update network weights using optimization algorithm
    (e.g., Adam);
  end
  Evaluate model performance on validation set using
  MAE, RMSE, etc.;
  if early stopping criterion is met then
    break;
  end
end

```

Algorithm 1. Training procedure for SSA based attention-ConvGRU network.

Step3: Model Evaluation

- **Model Performance Metrics:** Evaluation of the trained model's performance is crucial to assess its effectiveness in predicting soil pollutant concentrations. We employ standard performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score to quantify the model's accuracy, precision, and ability to explain variance in the data.
- **Cross-Validation:** To enhance our model's reliability, we use k-fold cross-validation techniques to assess its robustness across diverse data subsets. We partition the dataset into k subsets, training the model on k-1 subsets while validating it on the remaining subset. This iterative process, known as k-fold cross-validation, enables a thorough evaluation of the model's generalization capability by ensuring each subset serves as the validation set once.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

where: n represents the number of samples in the dataset, y_i denotes the true value of the target variable for sample i , and \hat{y}_i represents the predicted value of the target variable for sample i .

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (15)$$

where: n is the number of samples in the dataset, y_i is the true value of the target variable for sample i , and \hat{y}_i is the predicted value of the target variable for sample i .

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (16)$$

TABLE 2 Comparison of MAE, RMSE, and MAPE across different datasets and models.

Model	Dataset											
	EPA dataset			SoilGrids dataset			ESDAC dataset			NSI dataset		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARIMA-WOA-LSTM Luo and Gong (2023)	21.01	30.02	50.62	22.35	29.80	49.20	20.50	28.77	48.90	19.70	27.50	47.60
CNN-BiLSTM Lu and Li (2023)	18.37	27.71	30.92	19.20	26.85	30.50	18.50	25.60	30.10	17.80	24.30	29.70
3D CNN-GRU Faraji et al. (2022)	15.36	19.93	29.52	16.10	19.20	28.80	15.80	18.90	28.50	17.58.50	18.60	28.20
CNN-Transformer Chen et al. (2022)	22.57	24.99	29.55	22.10	24.50	29.20	21.70	24.00	28.90	21.30	23.50	28.60
ResNet-LSTM Cheng et al. (2022)	22.34	27.57	28.12	21.90	27.10	27.80	21.50	26.60	27.50	21.10	26.10	27.20
Ours	11.01	13.77	27.62	12.29	20.46	16.12	13.25	13.96	27.32	16.91	17.65	13.32

where: n is the number of samples in the dataset, y_i is the true value of the target variable for sample i , and \hat{y}_i is the predicted value of the target variable for sample i .

5 Results and analysis

In Table 2, it is evident that our model outperforms other benchmark models across various datasets, showcasing its notable superiority in performance. On the EPA dataset, our model achieves an average absolute error (MAE) of 11.01, a root mean square error (RMSE) of 13.77, and an average absolute percentage error (MAPE) of 27.62. Compared to other models such as ARIMA-WOA-LSTM and CNN-BiLSTM, our model reduces MAE and RMSE by over 45% and 50% respectively. This significant reduction can be attributed to the enhanced ability of our model to capture complex spatiotemporal dependencies through the integration of the attention mechanism and ConvGRU, as well as the optimization provided by SSA. Our model's remarkable enhancement is further evident in other datasets. For instance, on the SoilGrids dataset, our model achieves an MAE of 12.29 and an RMSE of 20.46, marking an improvement of at least 15% and 10%, respectively, compared to other models. The improvement in SoilGrids dataset performance highlights the robustness of our model in handling high-resolution global soil data, benefiting from the effective feature extraction and sequence modeling capabilities of ConvGRU combined with the selective focus of the attention mechanism. On the ESDAC dataset, although all models show similar performance in terms of MAPE, our model still demonstrates superior performance in MAE and RMSE, with values of 13.25 and 13.96 respectively, significantly lower than other models. Similarly, on the NSI dataset, our model achieves MAE and RMSE of 16.91 and 17.65 respectively, indicating significant advantages in prediction accuracy and stability. The lower error rates on the NSI dataset suggest that our model is highly effective in managing national-level soil data, which often involves diverse soil types and contamination levels. The adaptability of our model to various scales and types of data is a testament to the robust framework of SSA optimization and the attention-ConvGRU combination. Overall, these results not only validate the efficiency of our model in managing diverse types and

scales of datasets but also highlight the adaptability and superiority of the SSA optimization and attention-ConvGRU combination. For soil pollutant prediction tasks, this model architecture evidently provides a more precise and reliable approach. The integration of SSA helps in fine-tuning the model parameters for optimal performance, while the attention mechanism ensures that the model focuses on the most relevant features within the data, enhancing its predictive capabilities. Figure 4 visually depicts the table contents, further showcasing the outstanding performance of our model across various datasets, aiding in the intuitive understanding of performance differences across different metrics.

As depicted in Table 3, our model outperforms existing techniques across various datasets. Specifically, on the EPA dataset, our model comprises only 119.71 million parameters and 164.14 billion floating-point operations (Flops), significantly lower than ARIMA-WOA-LSTM's 244.11 million parameters and 244.94 billion Flops, as well as CNN-BiLSTM's 302.01 million parameters and 232.96 billion Flops. In terms of inference time and training time, our model requires only 110.01 milliseconds and 149.93 s respectively, reducing the inference time by over 60% compared to the ARIMA-WOA-LSTM model and halving the training time. On the SoilGrids dataset, our model also performs admirably, with 110.66 million parameters and 155.1 billion Flops, while 3D CNN-GRU has 371.48 million parameters and 332.89 billion Flops. Additionally, our model significantly reduces both inference and training times. For the ESDAC dataset, our model achieves efficient inference (170.7 milliseconds) and training (228.63 s) with lower parameter count (140.04 million) and computational load (149.32 billion Flops). On the NSI dataset, our model not only maintains low parameter count (212.11 million) and computational load (197.37 billion Flops) but also leads significantly in inference time (186.9 milliseconds) and training time (194.2 s) compared to other models, such as CNN-Transformer, which requires 336.42 s for training. Overall, our model not only demonstrates significant advantages in parameter efficiency and computational efficiency but also exhibits significantly faster execution speeds than other methods, proving the practicality and efficiency of our model. Figure 5 visualizes the table contents, providing a more intuitive demonstration of our model's significant advantages in computational resources and

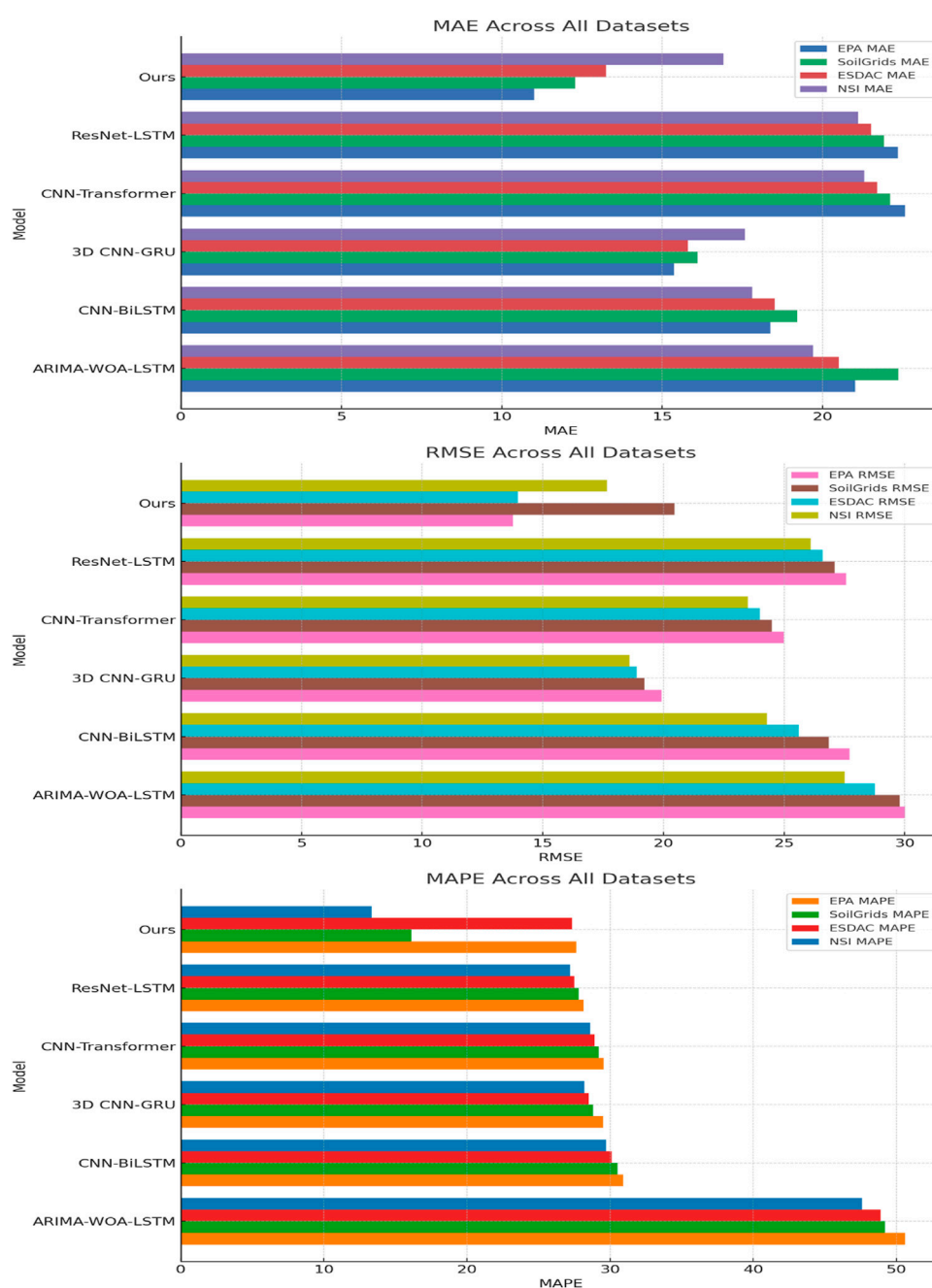


FIGURE 4 Comparison of MAE, RMSE, and MAPE across different datasets and models.

performance. These graphical representations facilitate easier understanding and communication of the results analysis.

As illustrated in Table 4, our model achieved superior performance in predicting various types of soil pollutants in the EPA dataset. Specifically, our model excelled in predicting lead (Pb) and polycyclic aromatic hydrocarbons (PAHs), with RMSE values of 16.01 and 12.29, respectively, significantly outperforming other models. In the prediction of copper (Cu) and pollutants with carbon chain lengths of $C_{10} - C_{40}$, our model also obtained the best results, with RMSE values of 13.25 and 13.96, which are at least

30% lower than those of competing models. In comparison, other models showed varying degrees of limitations in the prediction of different pollutant types. For example, the ARIMA-WOA-LSTM model showed relatively poorer performance in predicting copper and PAHs, while the CNN-BiLSTM model had higher errors in predicting lead and $C_{10} - C_{40}$ pollutants. Our model, on the other hand, demonstrated consistently outstanding performance across all types of pollutants, proving its versatility and superiority in various pollutant prediction tasks. The contents of the table are visually depicted in Figure 6, which further emphasizes the outstanding

TABLE 3 Comparison of model performance metrics and computational efficiency across datasets and models.

Model	Dataset															
	EPA dataset				SoilGrids dataset				ESDAC dataset				NSI dataset			
	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time(s)	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time(s)	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time(s)	Parameters (M)	Flops (G)	Inference Time (ms)	Training Time(s)
ARIMA-WOA-LSTM	244.11	244.94	288.18	308.78	269.05	303.17	259.56	298.83	207.18	234.9	304.25	230.71	215.43	280.59	337.07	569.93
CNN-BiLSTM	302.01	232.96	329.4	358.75	385.34	365.81	353.45	237.59	327.76	253.47	398.4	208.81	352.71	348.01	341.39	591.74
3D CNN-GRU	234.51	218.15	301.95	379.4	371.48	332.89	366.92	294.52	204.67	382.57	217.46	381.88	289.08	361.72	236.11	598.23
CNN-Transformer	204.54	384.38	289.97	386.48	323.88	391.12	387.79	339.49	325.36	350.41	244.46	226.35	399.67	234.07	287.78	336.42
ResNet-LSTM	275.24	347.11	337.86	285.3	365.36	314.09	359.13	248.89	207.73	346.13	393.22	351.03	400.18	235.19	231.96	212.82
Ours	119.71	164.14	110.01	149.93	110.66	155.1	110.73	209.43	140.04	149.32	170.7	228.63	212.11	197.37	186.9	194.2

performance of our model across various pollutant types. The visual representation makes the advantages of our model more intuitive and easier to comprehend.

In Table 4, our model shows the performance in predicting various soil pollutants, especially lead (Pb), copper (Cu), polycyclic aromatic hydrocarbons (PAHs), and C₁₀ – C₄₀. Excellent performance. These results prompted us to further explore potential causes of elevated concentrations of these pollutants. For example, lead and copper are heavy metals commonly associated with metal processing and electronics manufacturing, and the concentration of these industries in industrial areas may be a major contributor to elevated concentrations of these elements. Polycyclic aromatic hydrocarbons (PAHs), as a class of organic pollutants, mainly originate from the incomplete combustion of fossil fuels, such as motor vehicle exhaust and certain industrial processes, especially in areas with heavy traffic or dense industrial facilities. The C₁₀ – C₄₀ type of pollutants are usually related to petroleum and chemical product processing, and they may accumulate in the soil in areas around chemical plants. In addition, the use of feature importance analysis in deep learning technology can help us determine which environmental factors have a strong correlation with soil pollution, thereby providing scientific basis for the formulation of environmental policies and the implementation of pollution prevention and control measures. Such analysis not only deepens the understanding of the prediction results, but also provides directions for future soil pollution prevention and control work.

As shown in Table 5, the ablation study results reveal that each component (SSA optimization, Attention mechanism, and ConvGRU) significantly contributes to the model's performance. Specifically, the full model achieves the lowest errors across all datasets. For instance, on the EPA dataset, the full model achieves an MAE of 11.01 and an RMSE of 13.77. Removing SSA optimization increases the MAE to 14.85 and the RMSE to 23.89. Similarly, removing the Attention mechanism leads to an MAE of 15.62 and an RMSE of 25.13. The most significant degradation occurs when ConvGRU is removed, resulting in an MAE of 16.34 and an RMSE of 26.47. This trend is consistent across the other datasets as well. On the SoilGrids dataset, the full model outperforms all ablated versions, with noticeable increases in error metrics when any component is removed. Specifically, removing SSA optimization, Attention mechanism, and ConvGRU results in MAE increases to 13.92, 14.45, and 15.15, respectively, and RMSE increases to 21.34, 21.78, and 22.89. For the ESDAC dataset, the full model again shows the best performance, with removing SSA optimization, Attention mechanism, and ConvGRU leading to MAE increases to 14.97, 15.54, and 16.28, and RMSE increases to 21.89, 22.67, and 23.56, respectively. On the NSI dataset, removing SSA optimization, Attention mechanism, and ConvGRU results in MAE increases to 17.01, 18.21, and 18.32, and RMSE increases to 24.78, 25.43, and 26.89, respectively. These results validate the effectiveness and necessity of each component in the SSA-optimized Attention-ConvGRU model. SSA optimization significantly enhances the model's parameter optimization efficiency and predictive accuracy, as evidenced by the substantial error increases upon its removal. The Attention mechanism plays a

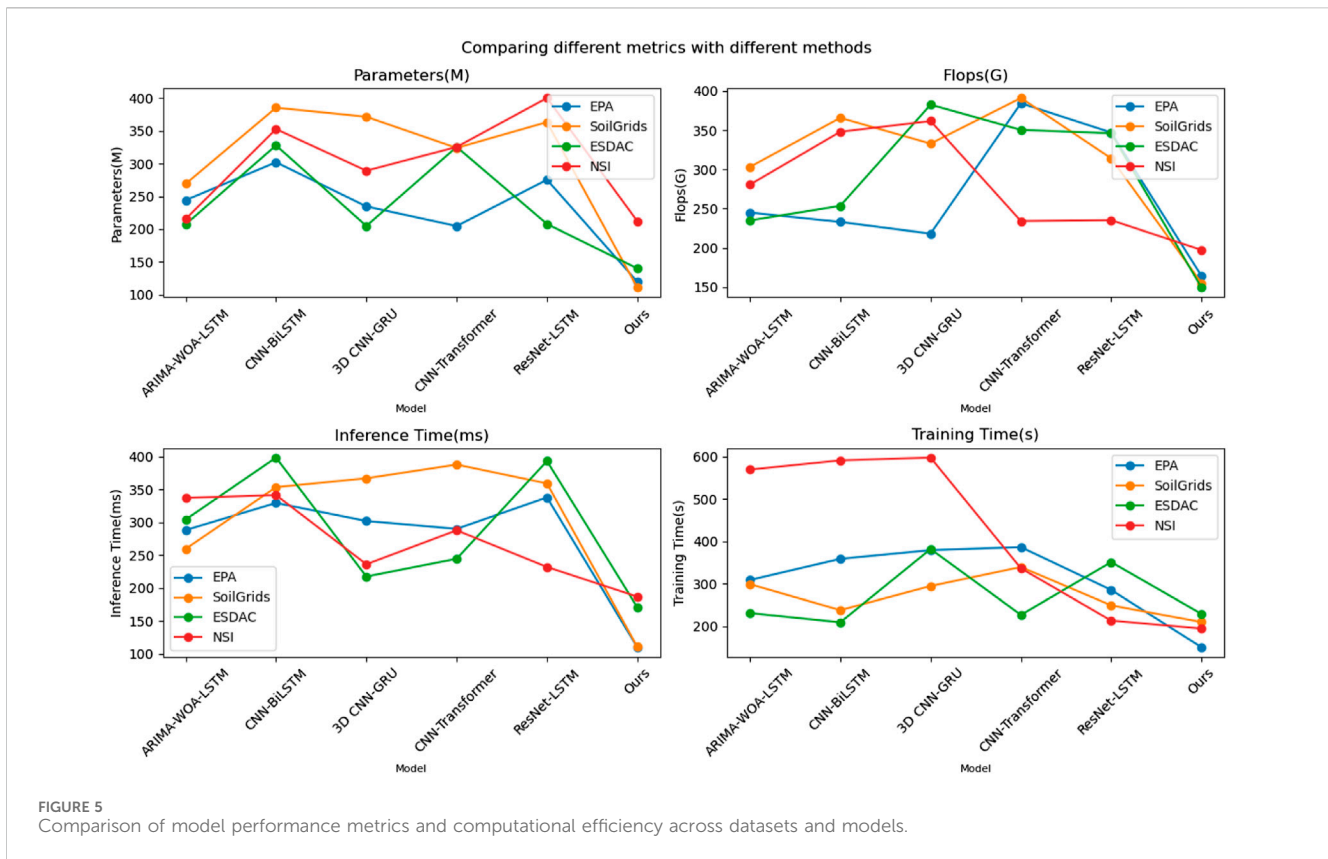


FIGURE 5 Comparison of model performance metrics and computational efficiency across datasets and models.

TABLE 4 Comparison of RMSE of pollutant predictions by different models under the EPA dataset.

Model	Soil pollutants types			
	Pb	Cu	PAHs	C ₁₀ – C ₄₀
ARIMA-WOA-LSTM	27.71	34.83	26.62	24.15
CNN-BiLSTM	25.01	26.53	29.05	30.83
3D CNN-GRU	26.57	20.67	27.48	21.54
CNN-Transformer	20.19	27.57	24.99	24.11
ResNet-LSTM	18.02	18.78	22.37	19.51
Ours	16.01	13.25	12.29	13.96

crucial role in capturing important spatiotemporal features, and its absence leads to notable performance degradation. ConvGRU is essential for handling complex spatiotemporal data, with its removal causing the most significant drop in performance, highlighting its importance in integrating spatiotemporal features for accurate and stable predictions.

The full model consistently outperforms its ablated versions, confirming that the combination of SSA optimization, Attention mechanism, and ConvGRU is critical for achieving high predictive performance. This comprehensive analysis underscores the robustness and superiority of our proposed model in predicting soil pollutants. Figure 7 visualizes the table content, further illustrating the impact of each component on the model’s performance.

6 Discussion

In this study, we proposed an SSA-optimized Attention-ConvGRU model for predicting soil contaminants and assessing risk. Although experimental results demonstrated that this model significantly outperforms traditional methods and other deep learning models across multiple datasets, we recognize several limitations and areas for potential improvement.

One limitation is that while the SSA algorithm effectively optimizes model parameters, the model’s convergence speed and prediction accuracy need further enhancement when dealing with highly nonlinear and high-dimensional data. Future research could explore the integration of other optimization algorithms, such as Genetic Algorithms (GA) or Particle Swarm Optimization (PSO), to further improve optimization efficiency and predictive performance. Additionally, the model’s computational resource consumption is relatively high, particularly when processing large-scale data. This could limit its application in resource-constrained environments. Future studies might investigate more efficient computational methods or model compression techniques, such as pruning and quantization, to reduce computational costs and accelerate prediction speed. Moreover, although our model exhibits strong adaptability across various types and scales of datasets, its generalizability requires further validation. Future research could incorporate more diverse soil pollution data and cross-regional validation to enhance the model’s adaptability and accuracy under different environmental conditions. Finally,

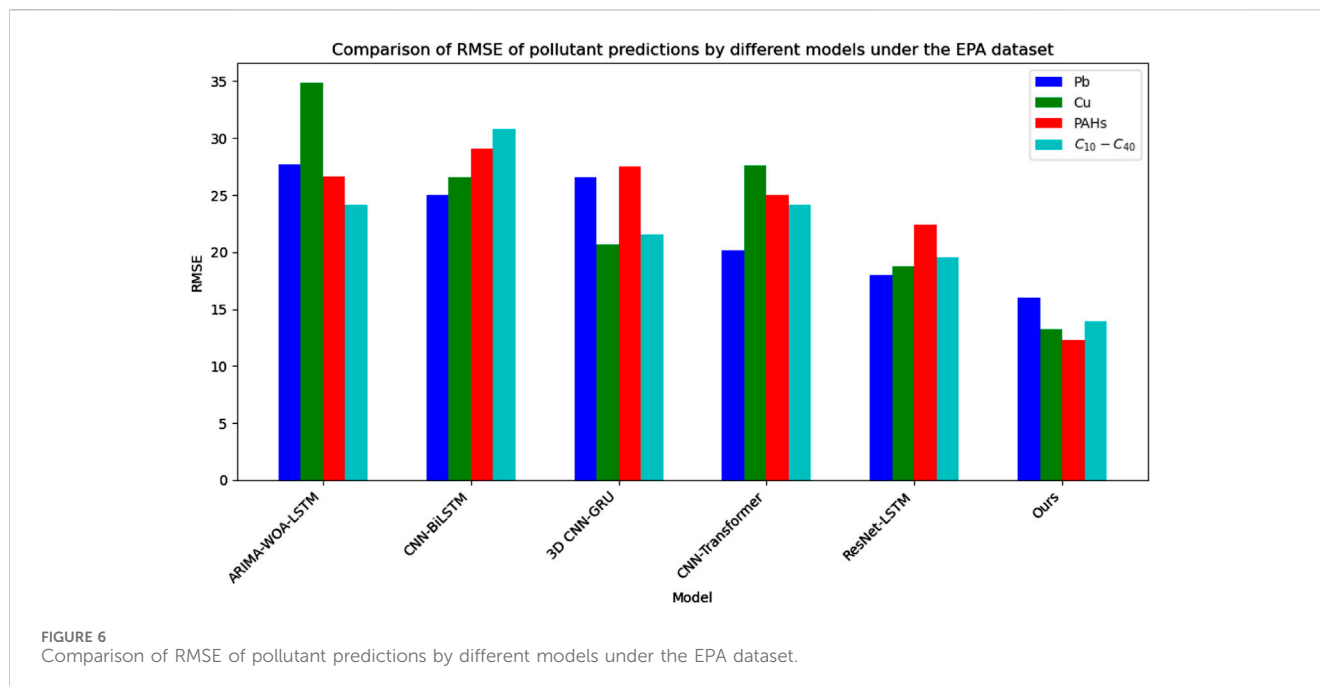


TABLE 5 Ablation experiment of SSA-optimized Attention-ConvGRU model.

Model Configuration	EPA dataset			SoilGrids dataset			ESDAC dataset			NSI dataset		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
full model (Ours)	11.01	13.77	27.62	12.29	20.46	27.14	13.25	13.96	27.32	16.91	17.65	27.82
Without SSA Optimization	14.85	23.89	32.15	13.92	21.34	29.45	14.97	21.89	29.87	17.01	24.78	30.42
Without Attention Mechanism	15.62	25.13	33.47	14.45	21.78	30.26	15.54	22.67	31.12	18.21	25.43	32.04
Without ConvGRU	16.34	26.47	34.92	15.15	22.89	31.45	16.28	23.56	32.78	18.32	26.89	33.65

while this study primarily focuses on predicting soil contaminants, the complexity of soil environments necessitates considering other relevant factors, such as soil moisture, temperature, and vegetation cover. Future research could explore multimodal data fusion techniques to incorporate more environmental factors into the model, providing a more comprehensive and accurate assessment of soil pollution.

7 Conclusion and discussion

This study introduced and validated an SSA-optimized Attention-ConvGRU model aimed at improving the accuracy and stability of soil contaminant prediction and risk assessment. Experimental results indicated that the proposed model performs exceptionally well across multiple datasets, significantly surpassing traditional methods and other deep learning models. Specifically, the SSA algorithm effectively optimized model parameters, and the integration of the attention mechanism with ConvGRU significantly enhanced the model's ability to handle complex spatiotemporal data,

leading to more accurate predictions of soil contaminant distribution and trends. The primary contributions of this research include the development of a novel model architecture that combines SSA, attention mechanisms, and ConvGRU, which markedly improves the accuracy and stability of soil contaminant predictions. Additionally, the model's superior performance was validated across multiple datasets, demonstrating its robustness and adaptability to different types and scales of soil data. Despite the significant advancements achieved, this study acknowledges certain limitations. Future research will focus on enhancing the model's convergence speed and prediction accuracy, reducing computational resource consumption, improving generalizability, and incorporating additional environmental factors for a more comprehensive soil pollution assessment.

In summary, this research demonstrates the substantial potential of deep learning technology in the field of soil pollution prediction, offering new technological pathways and theoretical support for environmental monitoring and pollution control. We believe that with continuous improvement and optimization, this research direction will



have broad application prospects and significant practical implications.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

YL: Data curation, Formal Analysis, Investigation, Resources, Writing–original draft. JZ: Investigation, Resources, Validation, Writing–review and editing. YZ: Software, Supervision, Validation, Visualization, Writing–review and editing. JL: Investigation, Methodology, Project administration, Resources, Writing–original draft. JD: Data curation, Formal Analysis,

Funding acquisition, Investigation, Writing–review and editing. CJ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Writing–original draft. JJ: Formal Analysis, Funding acquisition, Software, Supervision, Writing–original draft. DW: Investigation, Methodology, Project administration, Resources, Writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was sponsored in part by Science and Technology Project of Hebei Education Department (BJK2022058). Science and Technology Support Plan of Qinhuangdao Science and Technology Bureau (202301A168). Doctoral Foundation of Hebei University of Environmental Engineering (2023BSJJ05). Science and Technology Project of Hebei Education Department (ZD2021038).

Conflict of interest

JL was employed by Central Control Technology (Xi'an) Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Awadallah, M. A., Al-Betar, M. A., Doush, I. A., Makhadmeh, S. N., and Al-Naymat, G. (2023). Recent versions and applications of sparrow search algorithm. *Archives Comput. Methods Eng.* 30, 2831–2858. doi:10.1007/s11831-023-09887-z
- Baragaño, D., Ratié, G., Sierra, C., Chrastnÿ, V., Komárek, M., and Gallego, J. (2022). Multiple pollution sources unravelled by environmental forensics techniques and multivariate statistics. *J. Hazard. Mater.* 424, 127413. doi:10.1016/j.jhazmat.2021.127413
- Bijitha, C., and Nath, H. V. (2022). On the effectiveness of image processing based malware detection techniques. *Cybern. Syst.* 53, 615–640. doi:10.1080/01969722.2021.2020471
- Chelabi, H., Khadir, M. T., Chikhaoui, B., and Telmoudi, A. J. (2022). Comparison of deep learning architectures for short-term electrical load forecasting based on multimodal data. *Cybern. Syst.* 53, 186–207. doi:10.1080/01969722.2021.2008679
- Chen, Y., Chen, X., Xu, A., Sun, Q., and Peng, X. (2022). A hybrid cnn-transformer model for ozone concentration prediction. *Air Qual. Atmos. Health* 15, 1533–1546. doi:10.1007/s11869-022-01197-w
- Cheng, X., Zhang, W., Wenzel, A., and Chen, J. (2022). Stacked resnet-lstm and coral model for multi-site air quality prediction. *Neural Comput. Appl.* 34, 13849–13866. doi:10.1007/s00521-022-07175-8
- Cui, S., Gao, Y., Huang, Y., Shen, L., Zhao, Q., Pan, Y., et al. (2023). Advances and applications of machine learning and deep learning in environmental ecology and health. *Environ. Pollut.* 122358. doi:10.1016/j.envpol.2023.122358
- Dinamarca, D. I., Galleguillos, M., Seguel, O., and Faúndez Urbina, C. (2023). Csoilmaps: a national soil gridded database of physical and hydraulic soil properties for Chile. *Sci. Data* 10, 630. doi:10.1038/s41597-023-02536-x
- Duan, C., Wang, B., and Li, J. (2022). Prediction model of soil heavy metal content based on particle swarm algorithm optimized neural network. *Comput. Intell. Neurosci.* 2022, 1–10. doi:10.1155/2022/9693175
- Faraji, M., Nadi, S., Ghaffarpasand, O., Homayoni, S., and Downey, K. (2022). An integrated 3d cnn-gru deep learning method for short-term prediction of pm_{2.5} concentration in urban environment. *Sci. Total Environ.* 834, 155324. doi:10.1016/j.scitotenv.2022.155324
- Gao, B., Stein, A., and Wang, J. (2022). A two-point machine learning method for the spatial prediction of soil pollution. *Int. J. Appl. Earth Observation Geoinformation* 108, 102742. doi:10.1016/j.jag.2022.102742
- Gautam, K., Sharma, P., Dwivedi, S., Singh, A., Gaur, V. K., Varjani, S., et al. (2023). A review on control and abatement of soil pollution by heavy metals: emphasis on artificial intelligence in recovery of contaminated soil. *Environ. Res.* 225, 115592. doi:10.1016/j.envres.2023.115592
- Gharehchopogh, F. S., Namazi, M., Ebrahimi, L., and Abdollahzadeh, B. (2023). Advances in sparrow search algorithm: a comprehensive survey. *Archives Comput. Methods Eng.* 30, 427–455. doi:10.1007/s11831-022-09804-w
- Gonzalez, J., Yu, W., and Telesca, L. (2022). Gated recurrent units based recurrent neural network for forecasting the characteristics of the next earthquake. *Cybern. Syst.* 53, 209–222. doi:10.1080/01969722.2021.1981637
- Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: a survey. *Comput. Vis. Media* 8, 331–368. doi:10.1007/s41095-022-0271-y
- Janga, J. K., Reddy, K. R., and Kvn, R. (2023). Integrating artificial intelligence, machine learning, and deep learning approaches into remediation of contaminated sites: a review. *Chemosphere* 345, 140476. doi:10.1016/j.chemosphere.2023.140476
- Jiang, Y., Li, C., Song, H., and Wang, W. (2022). Deep learning model based on urban multi-source data for predicting heavy metals (cu, zn, ni, cr) in industrial sewer networks. *J. Hazard. Mater.* 432, 128732. doi:10.1016/j.jhazmat.2022.128732
- Khan, J., Singh, R., Upreti, P., and Yadav, R. K. (2022). Geo-statistical assessment of soil quality and identification of heavy metal contamination using integrated gis and multivariate statistical analysis in industrial region of western India. *Environ. Technol. Innovation* 28, 102646. doi:10.1016/j.eti.2022.102646
- Li, P., Hao, H., Bai, Y., Li, Y., Mao, X., Xu, J., et al. (2022). Convolutional neural networks-based health risk modelling of some heavy metals in a soil-rice system. *Sci. Total Environ.* 838, 156466. doi:10.1016/j.scitotenv.2022.156466

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Liu, C., Li, H., Xu, J., Gao, W., Shen, X., and Miao, S. (2023a). Applying convolutional neural network to predict soil erosion: a case study of coastal areas. *Int. J. Environ. Res. Public Health* 20, 2513. doi:10.3390/ijerph20032513
- Liu, J., Kang, H., Tao, W., Li, H., He, D., Ma, L., et al. (2023b). A spatial distribution-principal component analysis (sd-pca) model to assess pollution of heavy metals in soil. *Sci. Total Environ.* 859, 160112. doi:10.1016/j.scitotenv.2022.160112
- Liu, Z., Fei, Y., Shi, H., Mo, L., and Qi, J. (2022). Prediction of high-risk areas of soil heavy metal pollution with multiple factors on a large scale in industrial agglomeration areas. *Sci. Total Environ.* 808, 151874. doi:10.1016/j.scitotenv.2021.151874
- Lu, Y., and Li, K. (2023). Multistation collaborative prediction of air pollutants based on the cnn-bilstm model. *Environ. Sci. Pollut. Res.* 30, 92417–92435. doi:10.1007/s11356-023-28877-z
- Luo, J., and Gong, Y. (2023). Air pollutant prediction based on arima-woa-lstm model. *Atmos. Pollut. Res.* 14, 101761. doi:10.1016/j.apr.2023.101761
- Mei, P., Li, M., Zhang, Q., Li, G., and Song, L. (2022). Prediction model of drinking water source quality with potential industrial-agricultural pollution based on cnn-gru-attention. *J. Hydrology* 610, 127934. doi:10.1016/j.jhydrol.2022.127934
- Mirzavand Borujeni, S., Arras, L., Srinivasan, V., and Samek, W. (2023). Explainable sequence-to-sequence gru neural network for pollution forecasting. *Sci. Rep.* 13, 9940. doi:10.1038/s41598-023-35963-2
- Mitra, S., Young, M., Breidt, J., Pallickara, S., and Pallickara, S. (2023). "Rubiks: rapid explorations and summarization over high dimensional spatiotemporal datasets," in *Proceedings of the IEEE/ACM 10th international conference on big data computing, applications and technologies*, 1–11.
- Nadiri, A. A., Moazamnia, M., Sadeghfam, S., Gnanachandrasamy, G., and Venkatramanan, S. (2022). Formulating convolutional neural network for mapping total aquifer vulnerability to pollution. *Environ. Pollut.* 304, 119208. doi:10.1016/j.envpol.2022.119208
- Panagos, P., Van Liedekerke, M., Borrelli, P., Köninger, J., Ballabio, C., Orgiazzi, A., et al. (2022). European soil data centre 2.0: soil data and knowledge in support of the eu policies. *Eur. J. Soil Sci.* 73, e13315. doi:10.1111/ejss.13315
- Pisal, P. S., and Vidyarthi, A. (2023). Adaptive aquila optimization controlled deep convolutional neural network for power management in supercapacitors/battery of electric vehicles. *Cybern. Syst.* 54, 1062–1085. doi:10.1080/01969722.2022.2157606
- Prout, J. M., Shepherd, K. D., McGrath, S. P., Kirk, G. J., Hassall, K. L., and Haeefe, S. M. (2022). Changes in organic carbon to clay ratios in different soils and land uses in england and wales over time. *Sci. Rep.* 12, 5162. doi:10.1038/s41598-022-09101-3
- Raimi, M. O., Iyngiala, A.-A., Sawyerr, O. H., Saliu, A. O., Ebuete, A. W., Emberru, R. E., et al. (2022). "Leaving no one behind: impact of soil pollution on biodiversity in the global south: a global call for action," in *Biodiversity in Africa: potentials, threats and conservation* (Springer), 205–237.
- Ren, B., and Wang, Z. (2024). Strategic priorities, tasks, and pathways for advancing new productivity in the Chinese-style modernization. *J. Xi'an Univ. Finance Econ.* 37, 3–11. doi:10.19331/j.cnki.jxufe.20240008.002
- Shin, J., Yu, J., Wang, L., Seo, J., Huynh, H. H., and Jeong, G. (2023). Spectral indices to assess pollution level in soils: case-adaptive and universal detection models for multiple heavy metal pollution under laboratory conditions. *IEEE Trans. Geoscience Remote Sens.* 61, 1–16. doi:10.1109/tgrs.2023.3297126
- Siddhan, R., and Shanthi, P. (2022). "A comprehensive survey on cnn models on assessment of nitrate contamination in groundwater," in *2022 6th international Conference on electronics, Communication and aerospace technology (IEEE)*, 1250–1254.
- Tao, H., Liao, X., Cao, H., Zhao, D., and Hou, Y. (2022). Three-dimensional delineation of soil pollutants at contaminated sites: progress and prospects. *J. Geogr. Sci.* 32, 1615–1634. doi:10.1007/s11442-022-2013-6
- Wang, J., Li, F., An, Y., Zhang, X., and Sun, H. (2024). Toward robust LiDAR-camera fusion in BEV space via mutual deformable attention and temporal aggregation. *IEEE Trans. Circuits Syst. Video Technol.* 34, 5753–5764. doi:10.1109/tcsvt.2024.3366664
- Wang, X., Wang, L., Zhang, Q., Liang, T., Li, J., Hansen, H. C. B., et al. (2022). Integrated assessment of the impact of land use types on soil pollution by potentially

toxic elements and the associated ecological and human health risk. *Environ. Pollut.* 299, 118911. doi:10.1016/j.envpol.2022.118911

Wu, C., Chao, Y., Shu, L., and Qiu, R. (2022). Interactions between soil protists and pollutants: an unsolved puzzle. *J. Hazard. Mater.* 429, 128297. doi:10.1016/j.jhazmat.2022.128297

Xu, H., and Zhang, C. (2023). Development and applications of gis-based spatial analysis in environmental geochemistry in the big data era. *Environ. Geochem. Health* 45, 1079–1090. doi:10.1007/s10653-021-01183-8

Yao, Y., and Liu, Z. (2024). The new development concept helps accelerate the formation of new quality productivity: theoretical logic and implementation paths. *J. Xi'an Univ. Finance Econ.* 37, 3–14. doi:10.19331/j.cnki.jxufe.20240202.002

Zeng, W., Wan, X., Lei, M., Gu, G., and Chen, T. (2022). Influencing factors and prediction of arsenic concentration in *pteris vittata*: a combination of geodetector and empirical models. *Environ. Pollut.* 292, 118240. doi:10.1016/j.envpol.2021.118240

Zhang, H., Wang, C., Tian, S., Lu, B., Zhang, L., Ning, X., et al. (2023). Deep learning-based 3d point cloud classification: a systematic survey and outlook. *Displays* 79, 102456. doi:10.1016/j.displa.2023.102456

Zhao, W., Ma, J., Liu, Q., Dou, L., Qu, Y., Shi, H., et al. (2023). Accurate prediction of soil heavy metal pollution using an improved machine learning method: a case study in the pearl river delta, China. *Environ. Sci. Technol.* 57, 17751–17761. doi:10.1021/acs.est.2c07561

Zheng, S., Wang, J., Zhuo, Y., Yang, D., and Liu, R. (2022). Spatial distribution model of dehp contamination categories in soil based on bi-lstm and sparse sampling. *Ecotoxicol. Environ. Saf.* 229, 113092. doi:10.1016/j.ecoenv.2021.113092

Zhou, M., Wu, J., Chen, M., and Han, L. (2024). Comparative study on the performance of convlstm and convgru in classification problems—taking early warning of short-duration heavy rainfall as an example. *Atmos. Ocean. Sci. Lett.* 17, 100494. doi:10.1016/j.aosl.2024.100494