Multi-Source Heterogeneous Data-Driven Intelligent Prediction for Landslide Dam Longevity

Kang Peng  
Tianxing Ma  
Yang Chen  
Hengyu Liu  
Yun Lin  
Mingzhi Zhang  
Kun Luo  
Shijie Xie  
Xiangqi Hu  
Tianfeng Qi

Central South University  
Zhejiang University  
Central South University  
Central South University  
Central South University  
Central South University  
Central South University  
Southeast University  
Wuhan University  
Shanxi Huangyang Group New Energy Co.,Ltd

Article

Keywords: Unmanned Aerial Vehicle(UAV) 3D Laser Scanning, landslide longevity prediction, intelligent forecasting model, intelligent optimization algorithm, ensemble learning, SHAP method, Multi-source Heterogeneous Data-driven Approach

Posted Date: March 22nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3956793/v1
License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: There is NO Competing Interest.
Multi-Source Heterogeneous Data-Driven Intelligent Prediction for Landslide Dam Longevity

Tianxing Ma¹, Yang Chen², Kang Peng²*, Hengyu Liu²*, Yun Lin², Mingzhi Zhang², Kun Luo², Shijie Xie³, Xiangqi Hu⁴, Tianfeng Qi⁵

¹ Ocean College, Zhejiang University, Zhoushan 316021, Zhejiang, China
² School of Resources and Safety Engineering, Central South University, Changsha 410083, Hunan, China
³ School of Civil Engineering, Southeast University, Nanjing 210096, Jiangsu, China
⁴ School of Civil Engineering, Wuhan University, Wuhan 430062, Hubei, China
⁵ Shanxi Huayang Group New Energy Co., Ltd., Yangquan 045000, Shanxi, China

Abstract:

Accurate prediction of the longevity of Ldam, as temporary or permanent hydraulic structures, is crucial for ensuring downstream safety of lives and properties. This study proposes an intelligent forecasting model to accurately predict the longevity of Ldam. Firstly, a database containing a large amount of Ldam data is collected and organized, with key factors selected as indicators for the prediction system. Statistical indicators of the database are calculated, and in-depth analysis is conducted using correlation heatmaps and violin plots. Secondly, an intelligent forecasting model is constructed based on an improved intelligent optimization algorithm and ensemble learning. The model consists of multiple base learners (MLP, SVR, CatBoost) and a meta-learner (LightGBM). To effectively improve model performance, an improved intelligent optimization algorithm called IGTO is proposed to optimize the hyperparameters of the meta-learner within the stacked ensemble learning framework. In the analysis of the model's prediction results, prediction plots and regression plots are provided, and a series of evaluation metrics (R², Adj-R², RMSE, MAE, MAPE, VAF) are calculated. The results demonstrate that the IGTO stacked model exhibits high accuracy and reliability in predicting the longevity of Ldam, with computed values of R² = 0.98, Adj-R² = 0.98, RMSE = 93.98, MAE = 48.59, MAPE = 0.46, VAF = 98.29, showing good agreement with actual observed
values. Furthermore, the model outperforms other prediction models and previous empirical formulas, validating the effectiveness and practicality of the IGTO improved optimization algorithm and ensemble learning framework. Additionally, the SHAP method is employed to assess the importance and impact of each input parameter on the model's predictions, quantifying the significance of each influencing factor on Ldam longevity and providing reference for engineering professionals. Moreover, collaborations were conducted with technical personnel from a local enterprise in Yangquan City, Shanxi Province, where on-site investigations were carried out and first-hand data of 46 sets of Ldam were obtained using drone-based 3D laser scanning technology. These data further enhance the value of this study, enrich the currently scarce database of Ldam, and further validate the generalization ability of the model and its effectiveness in future practical engineering applications. In conclusion, this study enriches the currently scarce database of Ldam and provides an effective method for accurate longevity prediction, demonstrating significant practical significance.

**Keywords:** Unmanned Aerial Vehicle(UAV) 3D Laser Scanning, Ldam longevity prediction, intelligent forecasting model, intelligent optimization algorithm, ensemble learning, SHAP method, Multi-source Heterogeneous Data-driven Approach

1. **Introduction**

As a temporary or permanent hydraulic structure, landslide dams play a crucial role in water conservancy engineering and offer great convenience to human production and life. However, they also pose potential safety risks, among which the most significant is the failure and collapse of the dam body \(^1\). Once the dam body of a landslide dam is damaged, it will trigger catastrophic floods and water level rises, posing a severe threat to the people and property in downstream areas \(^2\text{-}^4\). The hazards of dam failure cannot be underestimated, as when the landslide dam is subjected to long-term scour and erosion by water flow or external factors such as earthquakes and debris flows, the dam body may rupture, slip, or even collapse,
leading to the sudden breach of the landslide dam lake or river \cite{5-6} and causing enormous losses to human life and property. Therefore, studying the longevity of landslide dams is of utmost importance \cite{7-9}.

![Field photo of landslide dam](image)

**Figure 1** Field photo of landslide dam

In recent years, an increasing number of scholars have conducted research on the longevity of landslide dams \cite{10-12}. By delving into the factors such as structure, materials, and environment of landslide dams, they aim to explore methods and techniques for longevity assessment to improve the accuracy and reliability of longevity prediction. Shen et al. \cite{13}, through analyzing the landslides of multiple landslide dams, found that the longevity of a landslide dam is related to several key factors, and rainfall-type landslide dams tend to have shorter longevities due to higher water content and larger inflow rates. Fan et al. \cite{14} constructed a global database containing hundreds of landslide dams worldwide and, after statistical analysis, determined that the longevity of a landslide dam is related to the formation mode of landslides, with those formed by rock slides having the longest longevity. Additionally, studies have found that if a landslide dam does not breach in its first year of formation, it becomes more stable due to self-stabilization. Yan et al. \cite{15} employed numerical simulation techniques to explore the causal mechanisms of landslide dams at a microscopic scale and determined through numerical modeling that higher friction coefficients and bond strengths increase the longevity of landslide dams.

After extensive research, it has been found that the structure and materials of
Landslide dams are affected by various factors over time, such as natural disasters like water erosion, corrosion, earthquakes, and landslides \cite{16-17}. Therefore, accurately estimating the longevity of landslide dams and promptly implementing corresponding maintenance measures is crucial for their normal operation and safety \cite{18}. Lai et al. \cite{19} conducted a comprehensive analysis of 65 landslide incidents in Japan and identified the key factors affecting the longevity of landslide dams to be the catchment area, dam height, slope height, dam width, and dam volume, with the catchment area having the greatest impact. They established a multiple regression model to predict the longevity of landslide dams, but the prediction results were not satisfactory, indicating the poor applicability of traditional statistical methods and empirical formulas for predicting the longevity of landslide dams. With the development of mathematical and statistical science and artificial intelligence technology, many scholars have begun to apply big data and machine learning techniques to the prediction of landslide dam longevity \cite{20}. Shi et al. \cite{21} divided the longevity of landslide dams into seven levels, ranging from less than an hour to more than a year, and used the k-nearest neighbor algorithm to fill in missing data. They successfully established a model for predicting the type and value of the longevity of landslide dams, with high prediction accuracy, and applied this prediction model to three landslide cases, successfully verifying its accuracy and providing a new model for predicting the longevity of landslide dams. Shen et al. \cite{13} divided the longevity of landslide dams into three stages and established models for predicting the longevity of each stage and the overall longevity. The prediction results showed that the three-stage longevity prediction model had better prediction performance due to its consideration of the influencing factors at each stage. These research results provide many methods for predicting the longevity of landslide dams, but they have not been able to overcome the problems of low accuracy and very limited applicability \cite{22-23}. In addition, the methods of longevity prediction of landslide dams are summarized in Table 1.

To address the limitations of the current research status, this study fully considers the impact of multiple key factors on the longevity of landslide dams and constructs a Stacking ensemble learning framework to predict the longevity of
landslide dams. Stacking ensemble learning is a machine learning method that combines the predictions of multiple base learners to improve overall prediction performance [24-26]. By combining the output results of multiple models, the final prediction results are more accurate and reliable, with better generalization performance. Ensemble learning algorithms have made significant research progress in the field of machine learning and have been widely applied in computer science, biomedicine, financial risk control, traffic management, and geological disaster prevention and control [27-33]. To further improve the performance and effectiveness of intelligent prediction models, this research has improved the GTO optimization algorithm (Artificial Gorilla Troops Optimizer, GTO) [34], proposed an adaptive weight-based GTO optimization algorithm (IGTO), and automatically optimized the hyperparameters of the meta-learners in the Stacking ensemble framework, resulting in a certain improvement in the performance of the prediction results compared to traditional GTO optimization algorithms [35-37]. In addition, the study used the SHAP method to analyze and explain the prediction results, quantifying the importance of each influencing factor on the longevity of landslide dams. In summary, the research provides references and insights for the accurate prediction of the longevity of landslide dams, providing scientific basis for disaster prevention and reduction and water resources management.

**Table 1** Rapid prediction statistical equations or models of LDam longevity

<table>
<thead>
<tr>
<th>Equations/Models (longevity)</th>
<th>Consideration factor</th>
<th>Note</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 10^{0.6847 \log V_d}$</td>
<td>$V_d$</td>
<td>Average longevity equation</td>
<td>[40]</td>
</tr>
<tr>
<td>$Y = 77.911 (V'_d / Q_s)^{0.5979}$</td>
<td>$V'_d$</td>
<td>Longevity Estimation Based on Overflow Time</td>
<td>[40]</td>
</tr>
<tr>
<td>$Y = -0.847 \log (Q_s) + 0.487 \log (H_d) + 2.215 \log (L_d) + 1.727 \log (W_d) - 2.556$</td>
<td>$Q_s, H_d, L_d, W_d$</td>
<td>34 cases from Japan</td>
<td>[41]</td>
</tr>
<tr>
<td>$Y = -0.438 \log (A'_w) + 0.678 \log (H_d) + 2.039 \log (L_d) + 1.973 \log (W_d) - 2.001$</td>
<td>$A'_w, H_d, L_d, W_d$</td>
<td>34 cases from Japan</td>
<td>[41]</td>
</tr>
<tr>
<td>$Y = -0.274 \log (A'_w) + 0.173 \log (W_d) + 0.864 \log (V'_d) + 0.322$</td>
<td>$A'_w, W_d, V'_d$</td>
<td>65 cases from Japan</td>
<td>[42]</td>
</tr>
</tbody>
</table>
\[ Y = -0.722 \log(A_{u}) - 0.438 \log(H_d) + 1.468 \log(H_s) + 1.016 \log(W_d) + 0.116 \log(V') + 1.746C_1 + 0.629C_2 - 1.059C_3 + 0.536C_4 + 6.165 \]

\[ \log(A_{u}, H_d, H_s, W_d, V') \]  
Same catchment and triggers \[ [42] \]

\[ Y = \left( \frac{H_d}{H_s} \right)^{0.082} \left( \frac{W_d}{H_d} \right)^{0.074} L_d \left( \frac{V'}{H_d} \right)^{0.054} \left( \frac{V'}{H_d} \right)^{2.184} \left( \frac{Q_{d}}{V'} \right)^{0.650} e^\beta \]

\[ H_d, L_d, Q_{d}, T_r \]  
Estimation of the full longevity \[ [13] \]

\[ Y = \left( \frac{H_d}{H_s} \right)^{3.285} \left( \frac{W_d}{H_d} \right)^{0.671} \left( \frac{L_d}{H_d} \right)^{1.436} \left( \frac{V'}{H_d} \right)^{0.309} \left( \frac{V'}{H_d} \right)^{2.188} \left( \frac{Q_{d}}{V'} \right)^{1.650} e^\beta \]

\[ H_d, L_d, Q_{d}, T_r \]  
Estimation of longevity in catchment stage \[ [13] \]

\[ Y = \left( \frac{H_d}{H_s} \right)^{3.285} \left( \frac{W_d}{H_d} \right)^{0.671} \left( \frac{L_d}{H_d} \right)^{1.436} \left( \frac{V'}{H_d} \right)^{0.309} \left( \frac{V'}{H_d} \right)^{2.188} \left( \frac{Q_{d}}{V'} \right)^{1.650} e^\beta \]

\[ H_d, L_d, Q_{d}, T_r \]  
Estimation of longevity in overcurrent stage \[ [13] \]

\[ Y = \left( \frac{H_d}{H_s} \right)^{0.771} \left( \frac{W_d}{H_d} \right)^{0.526} \left( \frac{L_d}{H_d} \right)^{0.211} \left( \frac{V'}{H_d} \right)^{0.647} \left( \frac{Q_{d}}{V'} \right)^{0.406} e^\beta \]

\[ H_d, L_d, Q_{d}, T_r \]  
Estimation of longevity in burst stage \[ [13] \]

\[ Y = \left( \frac{V'}{H_s} \right)^{0.333} \left( \frac{Q_{d}}{V'} \right)^{0.407} e^\beta \]

\[ V_d, H_s, Q_{d}, T \]
Estimation of the full longevity \[ [13] \]

\[ Y = -0.267 \log(V') + 2.204 \log(W_d) - 0.876 \log(B_s) - 1.68 \]

LightGBM

\[ W_d, V_d, B_s \]
Considering geometric factors \[ [39] \]

\[ H_d, L_d, W_d, V_d, V' \]
Missing data was processed \[ [21] \]

Notes: \( H_d \) is the LDam height (m); \( W_d \) is the LDam width (m); \( L_d \) is the LDam length (m); \( V' \) is the LDam volume (m³); \( V' \) is the dammed lake volume (m³); \( Q_{d} \) is the average annual discharge (m³/s); \( Q_{p} \) is the peak breach discharge (m³/s); \( H_d \) is the LDam lake depth (m); \( B_s \) is the backwater length (m); \( \alpha \) is Ldam material coefficient; \( \beta \) is Ldam trigger coefficient; \( A_{u} \) is the upstream catchment area (m²); \( H_s \) is the slope height of landslide (m); \( H_s \) is a constant(1m); \( T_r \) is a constant(1s).

2. Database analysis

2.1 Data source

This study collected 350 reliable and detailed case studies of landslide dams as a database (Shen\[^{13}\], Fan\[^{14}\], Lin\[^{38}\]). The database of 350 landslide dams was divided into training and testing sets, with the training set accounting for 80% of the total sample size and the testing set accounting for 20%. The dataset includes statistical data indicators for each sample, such as Dam height(m), Dam length(m), Dam width(m), Dam volume (10⁶m³), Dammed lake volume(10⁶m³), Upstream catchment area (km²), Triggers(earthquake/rainfall and snowmelt/others), Dam material(rock/debris/earth), Longevity (day) as shown in Table 2.
## Table 2 Statistical characteristics of the datasets

<table>
<thead>
<tr>
<th></th>
<th>STD</th>
<th>Kurt</th>
<th>Skew</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dam height (m)</td>
<td>80.40</td>
<td>56.86</td>
<td>5.80</td>
<td>1000</td>
<td>2</td>
<td>60.43</td>
<td>32</td>
<td>998</td>
</tr>
<tr>
<td>Dam length (m)</td>
<td>431.36</td>
<td>23.25</td>
<td>4.02</td>
<td>4000</td>
<td>5</td>
<td>404.69</td>
<td>250</td>
<td>3995</td>
</tr>
<tr>
<td>Dam width (m)</td>
<td>711.22</td>
<td>13.65</td>
<td>3.13</td>
<td>6000</td>
<td>20</td>
<td>657.97</td>
<td>362</td>
<td>5980</td>
</tr>
<tr>
<td>Dam volume (10^6 m^3)</td>
<td>201.28</td>
<td>146.13</td>
<td>11.14</td>
<td>3000</td>
<td>0.06</td>
<td>44.83</td>
<td>1.95</td>
<td>2999.99</td>
</tr>
<tr>
<td>Dammed lake volume (10^6 m^3)</td>
<td>175.13</td>
<td>23.31</td>
<td>4.35</td>
<td>1500</td>
<td>0.001</td>
<td>85.08</td>
<td>5</td>
<td>1499.99</td>
</tr>
<tr>
<td>Upstream catchment area (km^2)</td>
<td>20338.70</td>
<td>55.20</td>
<td>7.20</td>
<td>173484</td>
<td>0.5</td>
<td>3377.04</td>
<td>87.4</td>
<td>173483.5</td>
</tr>
<tr>
<td>Triggers(earthquake)</td>
<td>0.50</td>
<td>-1.97</td>
<td>0.21</td>
<td>1</td>
<td>0</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Triggers(rainfall and snowmelt)</td>
<td>0.50</td>
<td>-1.98</td>
<td>0.18</td>
<td>1</td>
<td>0</td>
<td>0.43</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Triggers(others)</td>
<td>0.29</td>
<td>5.81</td>
<td>2.79</td>
<td>1</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dam material(rock)</td>
<td>0.48</td>
<td>-1.71</td>
<td>0.55</td>
<td>1</td>
<td>0</td>
<td>0.38</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dam material(debris)</td>
<td>0.49</td>
<td>-1.87</td>
<td>0.37</td>
<td>1</td>
<td>0</td>
<td>0.41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dam material(earth)</td>
<td>0.42</td>
<td>-0.22</td>
<td>1.33</td>
<td>1</td>
<td>0</td>
<td>0.21</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Longevity (day)</td>
<td>678.83</td>
<td>26.70</td>
<td>4.73</td>
<td>5601</td>
<td>0.01</td>
<td>14474.79</td>
<td>23</td>
<td>5600.99</td>
</tr>
</tbody>
</table>

### 2.2 Database analysis

Based on the constructed database, the Pearson product-moment correlation coefficient (PPMCC) was utilized to quantify the degree of correlation between various influencing factor indicators. The PPMCC is a measure of the linear correlation between two variables and is calculated as follows:

\[
\rho_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]  

(1)

In the formula, \( X_i \) represents the value of the ith sample point for the influencing factor X, \( Y_i \) represents the value of the ith sample point for the influencing factor Y, \( \bar{X} \) denotes the mean of the influencing factor X, and \( \bar{Y} \) denotes the mean of the influencing factor Y.

A correlation heat map (Figure 2) was created using the dam database and the calculated correlations between the influencing factors. This heat map visually represents the relationships between the factors, with color intensity showing the strength and direction of correlations. It provides an overview of how the factors interact with each other in the context of Ldam, enhancing understanding of their
The correlation analysis based on the sample data of dam breaches is shown in Figure 2. The color of each square is proportional to the magnitude of the correlation coefficient, with darker colors indicating stronger correlations. Blue represents positive correlation while yellow represents negative correlation. From the plot, it can be observed that most of the indicators have weak correlations with each other, indicating their relative independence. The "earthquake" and "rainfall and snowmelt" inducing factors are significantly negatively correlated, consistent with their mutually exclusive relationship. The three influencing factors of "Dam material" also exhibit a certain degree of negative correlation, which is also consistent with their mutually exclusive relationship. The correlation coefficients between "Dam volume" and "Dam height", "Dam length", and "Dam width" are greater than 0.7, indicating a strong positive correlation. This suggests that as the size of the dam increases, so does its storage capacity. "Dammed lake volume" has strong positive correlations with "Dam height", "Dam length", "Dam width", and "Dam volume", with correlation...
coefficients all exceeding 0.6. This indicates that as the size and volume of the dam increase, the volume of the resulting reservoir also increases. There are also relatively strong positive correlations (greater than 0.5) between "Dam height", "Dam length", and "Dam width", suggesting that changes in these variables tend to occur together. Specifically, if the dam height increases, it is likely that the dam length and width will also increase, and the increase may be relatively large. The results of the correlation analysis of the influencing factors are consistent with reality, further verifying the scientific and reliable nature of the database.

Figure 3 Violin diagram of various variable

Figure 3 shows a violin plot based on various feature variables from the Ldam database, which provides a clear reflection of the data density distribution characteristics, as well as the data peaks and outliers, related to the longevity of the dams and their influencing factors. The plot indicates a wide range of dam sizes in the collected database, with the variation of each indicator exceeding 1000. This
comprehensive consideration of dams of different sizes increases the diversity and enhances the reliability of the model training and construction process. The data range for the upstream catchment area exceeds $1.7 \times 10^5$, indicating that the collected dam breach database covers dams from different watershed areas. This diversity is crucial for studying the longevity and influencing factors of dams since different sizes of watershed areas may have varying effects on the stability and retention capacity of dams. The six feature indicators of "Material and triggers of Ldam" are categorical variables following a discrete distribution, ranging between 0 and 1. This allows us to clearly observe the materials used in dam construction or the causes of dam failure. This information is equally important for studying the longevity and influencing factors of dams since different construction materials or reasons for failure may significantly affect the stability and retention capacity of dams. The data distribution of "Longevity of Ldam" is extremely wide, spanning from almost instantaneous instability and collapse to stable conditions lasting several years or even longer. Through further analysis of this data, we can better understand the longevity and stability of dams under different conditions.

3. Methods

3.1 Ensemble Learning Prediction Model

To enhance the accuracy and stability of predicting the longevity of a Ldam, this study constructed a predictive model by implementing a Stacking ensemble learning framework. Stacking ensemble learning effectively reduces the error of individual learners and enhances the generalization capability of the overall model by combining the predictions of multiple base learners. The study utilized MLP, SVR, and CatBoost algorithms as base learners, and employed LightGBM as the meta-learner. These algorithms exhibit strong performance and reliability in various scenarios, complementing each other's shortcomings and elevating the overall performance of the predictive model.

3.1.1 MLP
The multilayer perceptron (MLP), also known as a multilayered perception, is an artificial neural network capable of accurately predicting nonlinear data transformations. It consists of an input layer, a hidden layer, and an output layer, as illustrated in Figure 4.

Figure 4 General architecture of Multilayered perception

Figure 4 illustrates a multi-layer feedforward neural network, where neural nodes are represented by circles, and neurons are fully connected between layers without any intra-layer or cross-layer connections. The input layer receives inputs from external variables, and the hidden layer processes signals and transmits them to the output layer, where the output is generated by the output neurons. To introduce nonlinearity into the network and improve its generalization capability by approximating any form of nonlinear function, the activation function $f(x)$ is utilized. There are three types of activation functions, expressed as follows:

1. Sigmoid activation function:
Given that the Sigmoid and Tanh activation functions have high computational complexity and are prone to encountering the vanishing gradient problem, as they tend to approach saturation, leading to information loss when the transformation slows down and the derivative tends to zero, we chose the Relu activation function for this paper. The neurons in the first hidden layer are expressed using the twelve components of the input layer ($x_1$, $x_2$, $x_3$, ..., $x_{11}$, $x_{12}$).

$$z_j^1 = f \left[ \sum_{i=1}^{N_x} (\omega_{i,j,1}x_i + b_{i,j,1}) \right]$$  \hspace{1cm} (5)

where $z_j^1$ represents the $j$-th neuron on the 1st hidden layer, $N_x$ indicates the dimension of the input layer, which is 12 in this research, $x_i$ describes the $i$-th component of the input layer, $\omega_{i,j,1}$ represents the weight of the connection between the $i$-th component of the input layer and the $j$-th neuron on the 1st hidden layer, and $b_{i,j,1}$ denotes the bias.

During backpropagation, calculation results for each neuron on the 2nd hidden layer can be obtained recursively by applying Eq. (5) starting from the neurons on the 1st hidden layer:

$$z_j^2 = f \left[ \sum_{i=1}^{N_z} (\omega_{i,j,2}z_i^1 + b_{i,j,2}) \right]$$  \hspace{1cm} (6)

where $z_j^2$ represents the $j$-th neuron on the 2nd hidden layer, $N_z$ indicates the total number of neurons in the hidden layer, which is 6 in this paper, $x_i$ demotes the $i$-th component of the input layer, and $\omega_{i,j,2}$ describes the weight of the connection between the i-th component of the 1st hidden layer and the $j$-th neuron on the 2nd
hidden layer, and \( b_{i,j,2} \) represents the bias.

The expression of the output layer is as follows:

\[
y_j = f \left[ \sum_{i=1}^{N} \left( \omega_{i,j,M} z_i^M + b_{i,j,M} \right) \right]
\]

where \( y_j \) represents the \( j \)-th component of the output layer. In this paper, the prediction task involves multiple inputs and a single output, hence \( j \) is set to 1. \( M \) denotes the total number of hidden layers, which is set to 2.

### 3.1.2 SVR

Support Vector Regression (SVR) is a nonlinear regression algorithm that utilizes the principles of Support Vector Machines (SVM) by mapping the input space to a high-dimensional feature space for regression analysis\(^{(44)}\). SVR demonstrates excellent performance in handling nonlinear relationships and is suitable for capturing the complex influencing factors in predicting the longevity of Ldams.

![Schematic diagram of SVR principle](image)

**Figure 5** Schematic diagram of SVR principle

In this study, a given sample training set \( D_n \) is considered, where \( n \) represents the number of training samples, \( x \) denotes the independent variable in the training set, and \( y \) represents the corresponding dependent variable, which in this research refers to the
actual longevity of the debris dam. The nonlinear regression capability of Support Vector Machines (SVM) utilizes a nonlinear mapping function to transform the input data \( x \) into a high-dimensional linear space for regression analysis. Specifically, the regression function can be represented as follows:

\[
f(x) = \omega \cdot \phi(x) + b
\]  

(8)

In the formula, \( f(x) \) represents the regression hyperplane, where \( \omega \) and \( b \) are the coefficients of the regression hyperplane.

Introducing \( \varepsilon \)-insensitive loss function:

\[
L_\varepsilon(f(x), y) = \begin{cases} 
0, & |f(x) - y| < \varepsilon \\
|f(x) - y| - \varepsilon, & \text{others}
\end{cases}
\]  

(9)

In terms of controlling the fitting accuracy, the non-sensitive coefficient \( \varepsilon \) plays a crucial role. According to equation (9), if the difference between the predicted value \( f(x) \) and the actual value \( y \) is less than \( \varepsilon \), then \( f(x) \) is determined to have no loss; if the difference exceeds \( \varepsilon \), then it is considered that \( f(x) \) has a loss. Therefore, \( \varepsilon \) has a significant impact on the accuracy of regression analysis.

By introducing the penalty coefficient \( C \) and slack variable factor \( \xi_i, \xi_i^* \), the final optimization objective is:

\[
\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)
\]  

(10)

\[
\begin{align*}
&y_i - (\omega \cdot \phi(x_i) + b) \leq \varepsilon + \xi_i \\
&\xi_i, \xi_i^* \geq 0 (i = 1, 2, \cdots n)
\end{align*}
\]  

(11)

In the formula, \( \xi_i, \xi_i^* \) represents the slack variable for the \( i \)-th sample.

By utilizing Lagrange multipliers, the above function is transformed into its dual problem:

\[
\max \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \cdot K(x_i, y_j) - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*)
\]  

(12)
In the formula, $\alpha_i$, $\alpha_i^*$ is the Lagrange coefficient vector and $K(x_i, y_i)$ represents the kernel function.

By solving the dual problem, we obtain the final functional form of the Support Vector Machine (SVR):

$$ f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b $$

(14)

In the formula, $K(x_i, y_i)$ represents the kernel function. In this experiment, the radial basis kernel function RBF (Radial Basis Function) is used, which has the following functional form:

$$ K(x, x_i) = \exp \left( -\frac{\|x - x_i\|^2}{g^2} \right) $$

(15)

In the formula, $g$ represents the width of the radial basis kernel function.

### 3.1.3 CatBoost

CatBoost is an ensemble learning algorithm of gradient lifting decision tree (GBDT), which is highly scalable and accurate\[45\]. It can automatically deal with classification features and has strong anti-over-fitting ability, and is also widely used in regression prediction tasks. CatBoost is selected as one of the basic learners in order to make better use of the discrete characteristics such as inducing factors and dam materials in the dam database.

Firstly, CatBoost uses a weighted loss function to fit regression problems. For the $i$-th sample, its true value is $y_i$ and the model's predicted value is $f_i$. The loss function for this sample is:

$$ L_i(\omega) = \frac{1}{2} (y_i - f_i)^2 $$

(16)

In the formula, where $\omega$ represents the model's parameters.

To avoid overfitting, CatBoost also uses an $L_2$ regularization term, which is:
\[ R(\omega) = \frac{1}{2} \lambda \| \omega \|^2 \]  \hspace{1cm} (17)

In the formula, where \( \lambda \) is the regularization coefficient.

Therefore, the total loss function of CatBoost is:

\[ L(\omega) = \frac{1}{n} \sum_{i=1}^{n} \left[ L_i(\omega) + R(\omega) \right] \]  \hspace{1cm} (18)

In summary, CatBoost is an ensemble learning algorithm based on gradient boosting decision trees that uses a weighted loss function and L2 regularization term to fit regression problems, and uses the gradient descent algorithm to optimize the loss function.

3.1.4 LightGBM

LightGBM is a fast and efficient machine learning algorithm based on gradient boosting decision trees. It utilizes a histogram-based decision tree algorithm, resulting in lower memory consumption and faster training speed (Figure 6). In our study, we have chosen LightGBM as the meta-learner to integrate and optimize the predictions from various base learners, aiming to improve the overall performance of the model[46].

**Figure 6** Histogram algorithm of LightGBM

LightGBM incorporates the principles of gradient boosting algorithms, where in each iteration, a new decision tree is trained by fitting the residuals of the previous model. Assuming there have N samples and K features, denoted as \( X(i, j) \) for the j-th
feature of the i-th sample, with the corresponding actual value $Y(i)$. The regression prediction formula in LightGBM can be represented as:

$$Y(i) = \sum_{i=1}^{n} [\alpha \cdot h[X(i)]] + \epsilon$$  \hspace{2cm} (19)

In this formula, $\alpha$ is the weight coefficient of the decision tree, representing the contribution of each tree; $h[X(i)]$ is the predicted value of the i-th sample on each decision tree; $\epsilon$ is the error term of the model.

LightGBM optimizes the algorithm to minimize the error between the predicted values and the actual values, thereby obtaining the optimal decision tree weight coefficients and the error term of the model.

3.2 Intelligent optimization methods

To further enhance the predictive performance and capabilities of the model, the study opted to select and improve the Gorilla Troop Optimization (GTO) algorithm for multi-objective optimization of meta-learner hyperparameters. The GTO algorithm was introduced in 2021 by Abdollahzadeh et al., inspired by the social behavior of gorillas in their natural habitat [47].

3.2.1 Gorilla Troop Optimization (GTO)

The inspiration for the Gorilla Troop Optimization (GTO) algorithm stems from the social behavior of gorilla communities. The algorithm presents specific mathematical mechanisms to explain two stages of the optimization algorithm's exploration and exploitation, incorporating five behaviors observed in gorillas (Figure 7). This algorithm demonstrates strong performance and robustness in optimization problems.

During the exploration stage of the algorithm, three distinct behaviors are simulated and applied. Firstly, there is the migration to an unknown location to increase exploration of uncharted territories, aiding the algorithm in better navigating the problem space and identifying potential optimization solutions. Secondly, there is movement towards other gorillas, striking a balance between exploration and
exploitation. This behavior enables the algorithm to maintain a certain level of exploitation capability during the exploration process, preventing it from getting stuck in local optima. Finally, there is migration towards known locations, significantly enhancing the algorithm's ability to search different optimization spaces.

In the exploitation stage of the algorithm, two behaviors are simulated and applied. The first involves following a silverback gorilla, leveraging learning and imitation of excellent solutions to enhance its own performance. This behavior allows the algorithm to learn from identified optimization solutions and gradually improve and optimize itself. The second behavior involves competing for female gorillas, employing competition and selection to sift out the most outstanding solutions. This behavior enables the algorithm to select from multiple solutions and progressively develop even better solutions.

Figure 7 Development and Exploration Stage Schematic Diagram of the Gorilla Troop Optimization Algorithm

The optimization algorithm of artificial gorilla forces follows the following restrictions when searching for optimization:

(1) The optimization space of artificial gorilla troop optimization algorithm contains three types of solutions, X are gorilla position vector, and GX are the candidate position vectors of gorillas created at each stage. The silverback gorilla is
defined as the best solution;

(2) There is only one silverback gorilla in the whole gorilla population;

(3) Three types of solutions to the X, GX and silverback gorilla can accurately simulate the social life of gorillas in nature;

(4) Gorillas can increase their strength by finding better food sources. In this algorithm, the solution is created in each iteration and is called GX. In the iterative process, the found solution is the new (GX), and it will replace the current solution (X). On the contrary, it will be kept in (GX);

(5) Gorillas tend to live in groups, which makes them unable to live alone. Therefore, they will look for food together as a group and continue to live under the leadership of the silverback gorilla.

Next, the mathematical model of the GTO is established:

Firstly, the corresponding parameters of the algorithm are set, including population number \( N \), maximum iteration times \( \text{MaxIt} \), gorilla violence degree parameters \( \beta \), control parameters in development stage \( \omega \) and gorilla migration strategy parameters \( p \), and the gorilla population is initialized.

In the exploration stage, there are three mechanisms, namely, moving to unknown space, moving to known space and following other gorillas. These three mechanisms make the algorithm have superior space exploration ability. The first mechanism corresponds to the execution condition \( rand_1 < p \) that it is beneficial for gorillas to explore space randomly; The execution conditions of the second mechanism are \( rand_2 \geq 0.5 \) that it is beneficial to improve the search degree of the algorithm; The execution condition \( rand_3 < 0.5 \) corresponding to the third mechanism is to prevent the algorithm from deviating from the local optimum, which can be expressed by the following formula:

\[
GX(t+1) = \begin{cases} 
(UB - LB) \times r_1 + LB, & rand < p \\
(r_2 - C) \times X_r(t) + L \times H, & rand \geq 0.5 \\
X(t) - L \times (L \times (X(t) - GX_r(t)) + r_3 \times (X(t) - GX_r(t))), & rand < 0.5
\end{cases}
\]  

(20)

In the formula, \( GX(t+1) \) is location at the gorilla's next iteration; \( X(t) \) is current
position of the gorilla; \( r_1, r_2, r_3 \) and \( \text{rand} \) are random number, the range of values is \((0,1)\); \( UB \) is the upper bound of the variable; \( LB \) is the lower bound of the variable; \( x_\text{r} \) is one of the members of the randomly selected gorilla group in the whole population; The rest of the parameters \((C, L \text{ and } H)\) can be calculated by using the following formula:

\[
C = F \times \left( 1 - \frac{It}{MaxIt} \right)
\]

\[
F = \cos(2 \times r_4) + 1
\]

\[
L = C \times l
\]

\[
H = Z \times X(t)
\]

\[
Z = [-C, C]
\]

In the formula, \( C \) is current iteration value; \( MaxIt \) is the total number of iterations for performing the optimization operation; \( \cos \) is the cosine function; \( r_4 \) is random value, the range is \((0,1)\); \( F \) is iterative parameters, in the early stage of the optimization operation, it will produce a sudden change of value in a large interval, but this change interval decreases in the final stage; \( L \) is the cosine function, used to simulate the leadership of the silverback gorilla. In the real world, silverback gorilla may be unable to make the right decision to find food or control the group due to the lack of early experience of team leadership, but more experience can be gained in its leadership process, and when gained enough experience, it will show good stability; \( H \) is iterative parameters; \( Z \) is problem dimension of random values. At the end of the exploration phase, the cost of all the solutions was calculated, if \( GX(t) < X(t) \), the solution of \( GX(t) \) is then used as the solution of \( X(t) \).

In the artificial gorilla unit optimization algorithm, two behaviors were simulated during development: following the silverback gorilla and competing for the female gorilla. Both behaviors are switched by taking the value of \( C \) through.
If $C \geq w$, the gorilla will choose to follow the silverback gorilla by the following formula:

$$G_X(t+1) = L \times M \times (X(t) - X_{silverback}) + X(t)$$

(26)

$$M = \left( \frac{1}{N} \sum_{i=1}^{N} G_{X_i}(t) \right)^{\frac{1}{g}}$$

(27)

$$g = 2^L$$

(28)

In the formula, $X_{silverback}$ is optimum solution, corresponding to the silverback gorilla solution; $G_{X_i}(t)$ is position corresponding to each candidate gorilla at the iteration; $N$ is number of gorillas.

Conversely, the gorillas will choose to compete for the female gorilla, using the following formula:

$$G_X(t) = X_{silverback} - (X_{silverback} \times Q - X(t) \times Q) \times A$$

(29)

$$Q = 2 \times r_5 - 1$$

(30)

$$A = \beta \times E$$

(31)

$$E = \begin{cases} N_1, rand \geq 0.5 \\ N_2, rand < 0.5 \end{cases}$$

(32)

In the formula, $X_{silverback}$ is optimum solution, corresponding to the silverback gorilla solution; $Q$ is the intensity of the competition in the gorillas; $r_5$ is random number, and its value range is $(0,1)$; $A$ is the competitive strength coefficient for the gorilla; $E$ is impact of gorilla violence levels on solution dimensions; $rand$ is random number, and its value range is $(0,1)$.

If $rand \geq 0.5$, $E$ is normal distribution and random values in the dimension of the problem to be optimized; if $rand < 0.5$, $E$ is random values in a normal distribution.

In the development stage of the algorithm, the gorilla conducts social behavior in
a group form, calculating the cost of all solutions, if  $GX(t) < X(t)$, using the solution $GX(t)$ as the solution $X(t)$. Therefore, the optimal solution produced at this stage is also considered to be produced by the silverback gorilla.

### 3.2.2 Improved Gorilla Troop Optimization (IGTO)

To overcome the contradiction between global search and local exploitation in traditional artificial gorilla troop optimization (GTO) algorithms and prevent premature convergence, this study introduces adaptive weights to effectively expand the global search range and improve local exploitation capability.

The research proposes the introduction of two inertia weighting factors, one for the global search phase and the other for the local exploitation phase, to iteratively update the positions of individual gorillas in their respective neighborhoods. With a larger inertia weighting factor, the algorithm transitions to the global search phase, while with a smaller inertia weighting factor, it transitions to the local exploitation phase.

The adaptive weight during the global search phase is defined as follows:

$$k_1 = \sin\left(\frac{\pi t}{2M} + \pi\right) + 4$$  \hspace{1cm} (33)

In the formula, $k_1$ is adaptive weight.

Individual gorilla location update:

$$GX'_i = k_1X_i$$  \hspace{1cm} (34)

In the formula, $GX'_i$ is renewal location of the $i$th individual gorilla.

The adaptive weights for performing the local development are:

$$k_2 = \sin\left(\frac{\pi t}{2M} + \pi\right) + 2$$  \hspace{1cm} (35)

In the formula, $k_2$ is adaptive weight.

Individual gorilla location update:

$$GX'_i = k_2X_i$$  \hspace{1cm} (36)
In the formula, $GX_i$ id renewal location of the i-th individual gorilla.

3.3 IGTO-Stacking model for predicting the longevity of Ldam

An efficient combination of MLP, SVR, CatBoost, and LightGBM as base learners and meta-learner has been employed in this study to construct a powerful and stable integrated learning prediction model for the service longevity of Ldams, considering multiple influencing factors. This model will enable more accurate predictions of Ldam longevity under different conditions, providing more targeted recommendations and decision support for design and management. To achieve this, 80% of the Ldam sample data from the database was used to train the integrated learning prediction model, with the remaining 20% forming the test set to evaluate the model's predictive performance. Furthermore, in order to further enhance the performance of this integrated learning prediction model, the improved IGTO algorithm was utilized to automatically optimize the hyperparameters of the meta-learner.

The structure of the IGTO-based integrated learning prediction model for predicting Ldam longevity is illustrated in Figure 8.

![Figure 8 Workflow of IGTO-stacking Forecasting Model](image)
4. Prediction results of the model and analysis

In order to test the validity of the intelligent prediction model for predicting the longevity of the Ldam, the longevity of the Ldam samples was predicted, and compared with the prediction results of SVR, MLP, CatBoost, LightGBM\cite{21}, Stacking and GTO-Stacking models, so as to reasonably evaluate the performance and improvement effect of the IGTO-stacking prediction model compared with other models. Among them, the empirical formula method, due to the availability of data acquisition, selects several formulas that can obtain the calculation results of Ldam longevity\cite{40-42}, and the empirical formulas are as follow.

Empirical formula (I):

\[
Y = 10^{0.0417(\log Y)^{2.0657}}
\]  (37)

Empirical formula (II):

\[
Y = -0.438\log(A_w) + 0.678\log(H_d) + 2.039\log(L_d) + 1.973\log(W_d) - 2.001
\]  (38)

Empirical formula (III):

\[
Y = -0.274\log(A_w) + 0.173\log(W_d) + 0.864\log(V_d) + 0.322
\]  (39)

4.1 Model prediction results

According to the above empirical formulas (37), (38) and (39), the sample data of the Ldam are calculated, and the comparison between the calculated results and the actual longevity of the Ldam is shown in Figure 9:
The scatter plots presented in this study aim to compare the predicted longevity of a Ldam using three empirical formulas (Formulas 1, 2, and 3) with the actual longevity observed. The data used for this analysis covers a wide range of time durations: 0-100 days, 0-2000 days, 0-100000 days, and 0-1000000 days.

The results indicate a significant deviation between the predicted and actual longevity of the Ldam. In particular, the scatter plots for Formulas 2 and 3 show that the predicted longevity are clustered within the range of 0-10 days, whereas the actual longevity span a range of 0-10000 days. This suggests that the empirical formulas used in this study may not accurately estimate the longevity of the Ldam. On the other
hand, Formula 1 predicts a wider range of longevity, ranging from 0 to 800000 days. However, there still appears to be a notable discrepancy between the predicted and actual longevity.

The lack of correlation in the trend between the predicted and actual longevity further supports the notion that these empirical formulas may not account for all relevant factors that influence Ldam longevity. It is possible that other factors, such as the quality of construction materials or environmental conditions, may play a significant role in determining the longevity of the Ldam. Based on these findings, it is clear that further research and analysis are necessary to identify more accurate predictors of Ldam longevity.

In conclusion, the scatter plots demonstrate the disparities between the predicted longevity obtained from empirical formulas and the actual observed longevity of the Ldam. The results highlight the need for improved models or factors to be considered when predicting Ldam longevity accurately.

In this paper, MLP, SVR, Catboost, Stacking, GTO-Stacking and IGTO-Stacking models are used to predict the longevity of the Ldam samples, and the predicted results are shown in Figure 10. The predicted results of IGTO-Stacking model are highly consistent with the actual situation, and the fitting coefficient reaches 0.98. Compared with the Stacking model without parameter optimization and the GTO-Stacking model without improving GTO, the accuracy of the prediction results of IGTO-Stacking model has been greatly improved. However, the prediction performance of SVR, MLP and CatBoost models is inferior to that of Stacking model.
Figure 10 Regression diagram of prediction results of multiple models

(a) IGTO-Stacking
(b) GTO-Stacking
(c) Stacking
(d) SVR
(e) MLP
(f) CatBoost
4.2 Model prediction performance evaluation and comparison

According to the prediction results of several models, the self-prediction performance is quantitatively evaluated, and six statistical indicators are selected to evaluate the comprehensive performance of each model, including correlation coefficient ($R^2$), adjusted correlation coefficient (Adj.$R^2$), average absolute percentage error (MAPE), average absolute error (MAE), root mean square error (RMSE) and variance interpretation ratio (VAF). These indicators can comprehensively and objectively reflect the generalization ability of the model. The following are the mathematical definitions of these indicators (40)–(45):

\[
R^2 = 1 - \frac{\sum_{i=1}^{N} (y_{i,\text{mea}} - y_{i,\text{pre}})^2}{\sum_{i=1}^{N} (y_{i,\text{mea}} - y_{i,\text{mea}})^2}
\]  
(40)

\[
\text{Adj}.R^2 = 1 - \left[ (1 - R^2) \times \frac{N - 1}{N - k - 1} \right]
\]  
(41)

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{i,\text{mea}} - y_{i,\text{pre}})^2}
\]  
(42)

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_{i,\text{mea}} - y_{i,\text{pre}}|
\]  
(43)

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_{i,\text{mea}} - y_{i,\text{pre}}}{y_{i,\text{mea}}} \right| \times 100\%
\]  
(44)

\[
\text{VAF} = \left[ 1 - \frac{\text{var}(y_{i,\text{mea}} - y_{i,\text{pre}})}{\text{var}(y_{i,\text{mea}})} \right] \times 100\%
\]  
(45)

These statistical metrics were used to measure the performance of the model and were measured and compared in various aspects. In the formula, $y_{i,\text{mea}}$ and $y_{i,\text{pre}}$ indicate the measurements and the predictions, respectively, $y_{i,\text{mea}}$ represents $y_{i,\text{mea}}$ mean; $N$ represents the number of samples; $k$ represents the number of independent variables; The VAF is used to measure the difference between the measured and predicted outcomes. The ideal values of each evaluation index are shown in Table 3.
Table 3 The statistical metrics and corresponding ideal values

<table>
<thead>
<tr>
<th>Statistical metrics</th>
<th>Ideal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>0</td>
</tr>
<tr>
<td>MAE</td>
<td>0</td>
</tr>
<tr>
<td>RMSE</td>
<td>0</td>
</tr>
<tr>
<td>PI</td>
<td>2</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>1</td>
</tr>
<tr>
<td>$R^2$</td>
<td>1</td>
</tr>
<tr>
<td>VAF</td>
<td>100(%)</td>
</tr>
</tbody>
</table>

Based on equations (40)-(45), the predictive performance of the models constructed in this study, as well as existing empirical formulas\cite{37}\cite{38}\cite{39} and the LightGBM model\cite{21}, were quantitatively evaluated. The calculation results are presented in Table 4.

Table 4 Error comparison of models based on different imputation methods

<table>
<thead>
<tr>
<th>Models</th>
<th>R²</th>
<th>Adj.R²</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
<th>VAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGTO-Stacking</td>
<td>0.99</td>
<td>0.99</td>
<td>1.02</td>
<td>0.42</td>
<td>1.26</td>
<td>99.67</td>
</tr>
<tr>
<td>GTO-Stacking</td>
<td>0.99</td>
<td>0.99</td>
<td>39.52</td>
<td>14.21</td>
<td>27.53</td>
<td>99.57</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.99</td>
<td>0.99</td>
<td>98.60</td>
<td>45.48</td>
<td>22.18</td>
<td>99.21</td>
</tr>
<tr>
<td>CatBoost</td>
<td>0.97</td>
<td>0.97</td>
<td>141.13</td>
<td>51.93</td>
<td>97.09</td>
<td>97.82</td>
</tr>
<tr>
<td>MLP</td>
<td>0.97</td>
<td>0.97</td>
<td>176.05</td>
<td>36.80</td>
<td>24.75</td>
<td>97.68</td>
</tr>
<tr>
<td>SVR</td>
<td>0.96</td>
<td>0.96</td>
<td>178.92</td>
<td>40.08</td>
<td>48.92</td>
<td>96.42</td>
</tr>
<tr>
<td>Training sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k-Nearest NeighborLightGBM</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
<td>9.13</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>MCMC-LightGBM</td>
<td>0.89</td>
<td>-</td>
<td>-</td>
<td>14.19</td>
<td>33</td>
<td>-</td>
</tr>
<tr>
<td>Mean-LightGBM</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
<td>10.78</td>
<td>29</td>
<td>-</td>
</tr>
<tr>
<td>Empirical formula (I)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>0.01</td>
</tr>
<tr>
<td>Empirical formula (II)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>642.39</td>
<td>219.62</td>
<td>&gt;1000</td>
<td>0.04</td>
</tr>
<tr>
<td>Empirical formula (III)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>643.24</td>
<td>220.55</td>
<td>&gt;1000</td>
<td>0.03</td>
</tr>
<tr>
<td>Test sets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGTO-Stacking</td>
<td>0.98</td>
<td>0.98</td>
<td>93.98</td>
<td>48.59</td>
<td>0.46</td>
<td>98.29</td>
</tr>
<tr>
<td>GTO-Stacking</td>
<td>0.95</td>
<td>0.94</td>
<td>191.45</td>
<td>78.6</td>
<td>0.74</td>
<td>95.59</td>
</tr>
<tr>
<td>Stacking</td>
<td>0.91</td>
<td>0.89</td>
<td>272.33</td>
<td>151.01</td>
<td>0.79</td>
<td>90.88</td>
</tr>
<tr>
<td>CatBoost</td>
<td>0.88</td>
<td>0.85</td>
<td>315.62</td>
<td>169.44</td>
<td>11.82</td>
<td>87.78</td>
</tr>
<tr>
<td>MLP</td>
<td>0.86</td>
<td>0.83</td>
<td>339.48</td>
<td>144.95</td>
<td>42.32</td>
<td>86.32</td>
</tr>
<tr>
<td>SVR</td>
<td>0.83</td>
<td>0.80</td>
<td>366.23</td>
<td>187.57</td>
<td>7.86</td>
<td>83.44</td>
</tr>
<tr>
<td>k-Nearest NeighborLightGBM</td>
<td>0.95</td>
<td>-</td>
<td>-</td>
<td>11.04</td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td>Model Type</td>
<td>R²</td>
<td>Adj.R²</td>
<td>VAF(%)</td>
<td>MAE</td>
<td>RMSE</td>
<td>MAPE</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----</td>
<td>--------</td>
<td>--------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>MCMC-LightGBM</td>
<td>0.86</td>
<td>0.96</td>
<td>&lt;0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean-LightGBM</td>
<td>0.71</td>
<td>0.92</td>
<td>&lt;0.2</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
</tr>
<tr>
<td>Empirical formula (I)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Empirical formula (II)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>981.78</td>
<td>398.75</td>
<td>&gt;1000</td>
<td>0.04</td>
</tr>
<tr>
<td>Empirical formula (III)</td>
<td>&lt;0.2</td>
<td>&lt;0.2</td>
<td>982.88</td>
<td>400.15</td>
<td>&gt;1000</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Draw the radar drawing based on the calculation results of the prediction performance of each model in Table 4, see Figure 11 and Figure 12:

**Figure 11** Radar chart of prediction effect training set of multiple intelligent prediction models
Based on the data in Table 4 and the radar charts (Figure 11 and Figure 12), we can conduct a detailed analysis of the performance of each intelligent prediction model. Firstly, by observing the data in Table 4, it can be seen that the IGTO-Stacking model performs exceptionally well on both the training and testing sets, achieving good values in various indicators, especially $R^2$ and Adj.$R^2$, which are close to the ideal value of 1. This indicates that the model performs excellently in fitting the data and has good predictive ability, demonstrating strong generalization capability. In comparison, the GTO-Stacking and Stacking models perform well on the training set but exhibit larger errors on the testing set, particularly in terms of RMSE and MAE, suggesting that their generalization capabilities need to be improved and their predictive performance is unstable. Additionally, the CatBoost model shows quite good performance on both the training and testing sets, although it slightly lags behind the IGTO-Stacking, GTO-Stacking, and Stacking models, it still achieves high scores, indicating good generalization capability. On the other hand, some models based on empirical formulas and the LightGBM model generally perform poorly, especially on the testing set. This suggests that these models have larger errors when dealing with new data and weaker generalization capabilities.

In conclusion, the IGTO-Stacking model demonstrates the best performance in this study, achieving good predictive results on both the training and testing sets, and exhibiting strong generalization capability. However, other models show varying degrees of overfitting or insufficient generalization ability, which need further optimization and improvement.

In summary, this study comprehensively evaluates the performance of multiple intelligent prediction models and finds that the IGTO-Stacking model stands out with excellent predictive performance and high application potential. This provides important reference for prediction and decision-making in relevant fields. Future research will focus on improving and optimizing the models to further enhance prediction accuracy and generalization capability.

4.3 Feature contribution analysis based on SHAP method
The SHapley Additive exPlanations (SHAP) method is a powerful tool for feature contribution analysis in machine learning models. The method is based on the Shapley value, which is a concept from cooperative game theory that assigns a fair share of the total payout to each player in a game. In the context of machine learning, the Shapley value can be interpreted as the contribution of each feature to the prediction made by a model\(^{[48]}\).

The SHAP method provides a unified framework for computing the Shapley values in a computationally efficient manner. The method works by sampling subsets of features and computing the difference between the output of the model with and without each feature. This difference is then used to assign a weight to each feature, which represents its contribution to the prediction made by the model\(^{[49]}\).

One advantage of the SHAP method is its ability to handle complex models such as neural network models. The method can be applied to any model that produces a real-valued output, including models with non-linear interactions between features.

The study SHAP method was used to conduct an in-depth analysis of the model to explain the influence or importance of each feature to the longevity prediction of the Ldam. The analysis results are shown in Figure 13:

![Figure 13](image)

**Figure 13** SHAP analysis of various influencing factors

According to the SHAP method, figure 14 was drawn to show the contribution of influencing factors to the longevity of the Ldam.
By observing Figure 14, it can be seen that each feature has a relative importance in predicting the longevity of a Ldam. In the figure, each rectangle represents a feature, and the height of the rectangle indicates the degree of influence of the feature on the prediction results, which has already been quantified in the figure. If the rectangle is higher, it means that the corresponding value of the feature is higher and contributes more to the prediction results; if the rectangle is lower, it means that the corresponding value of the feature is lower and contributes less to the prediction results.

According to the figure, the importance of dam width is the highest among all influencing factors, reaching 0.341, indicating that it has the greatest impact on the longevity of the Ldam. The next most important influencing factors are upstream catchment area, dam length, dam volume, dam height, and dam volume, with importance values exceeding 0.1, respectively 0.177, 0.115, 0.104, 0.081, and 0.073.
indicating that these five influencing factors also play a key role in the change of the Ldam longevity.

The remaining six influencing factors have an importance level below 0.1, and these six influencing factors all belong to the inducing factors of the trigger factor and the Ldam material, with a proportion of less than 0.05, which can be considered as relatively minor factors.

In addition, the study also observed the interaction between different features. In the figure, the size and scale of the Ldam and the environmental factors such as the upstream catchment area seem to have similar and relatively important contribution levels, which may indicate that there is some kind of joint action between them.

In summary, engineering technicians should pay special attention to the six most important influencing factors, namely dam width, upstream catchment area, dam length, dammed lake volume, dam height, and dam volume. Their importance ranks ahead of all other influencing factors, and the importance values of individual influencing factors are all greater than 0.05. The importance level of these six influencing factors accounts for more than 89% of all influencing factors. Although the material composition of the Ldam and the trigger factor of Ldam only account for about 11%, their importance cannot be ignored. The quality and characteristics of the Ldam composition may have a significant impact on its longevity, while the trigger of the Ldam may have a significant impact on the stability and safety of the Ldam. Therefore, in designing, constructing, and maintaining the Ldam, engineering technicians should focus on the size and scale of the Ldam and ensure that it is within a safe and reasonable range. At the same time, sufficient research and evaluation should also be conducted on the composition of the Ldam and the inducing factors of the Ldam formation to ensure the safety and reliability of the Ldam.

5. Engineering case application

5.1 Project cases
Case samples of the Ldam project were collected in Qiansuohuang Village and Lililintou Village in Pingding County, Shanxi Province, and in hedi village, Hongtuyan Village, Guzhuang Mine Community, Yankan Village and Shandi Village in the suburbs of Yangquan City, Shanxi Province, with a total of 61 groups (Appendix), which added rich data to the dam database which is very scarce in current research.

The data collection method uses the Leica BLK2FLY autonomous flight 3 D laser scanner to measure the research object efficiently and accurately. Through the equipment (Figure 15), the high-resolution point cloud data of the Ldam and the surrounding terrain of the dammed lake can be obtained in a short time. These data can be used to generate detailed digital terrain models, and can provide accurate measurement information of Ldams and their formed dammed lakes. Information about the remaining influencing factors of the Ldam is obtained through the engineering and technical personnel of Huayang new material technology group co.,ltd.
5.2 Application situation of the intelligent prediction models

The longevity of the engineering case sample is predicted according to the existing empirical formula(I)(II)(III) and the research-constructed intelligent prediction model, respectively. The prediction results are shown in Figure 16 and Figure 17.

Figure 15 Three-dimensional laser scanning of Ldam and its surrounding environment

Figure 15 Longevity prediction results of Ldam
Based on the analysis of Figures 15 and 16, it is evident that the IGTO-Stacking model developed in this study exhibits the best predictive performance with a high coefficient of determination (R-squared) value of 0.97. This indicates that the intelligent predictive model constructed in this research is effective for practical engineering applications. Furthermore, other ensemble learning-based predictive models, such as GTO-Stacking and Stacking, also demonstrate relatively high accuracy and generalization ability, although slightly inferior to the IGTO-Stacking model. The CatBoost, MLP, and SVR models exhibit a gradual decrease in accuracy, but their predictions still fall within an acceptable range. On the other hand, it is apparent that traditional empirical formulas have limited generalization ability and poor applicability to new engineering cases, resulting in significant errors.

Therefore, in practical engineering application, it is suggested to use the intelligent prediction model to predict the Ldam. In practice, the accuracy and generalization ability of the prediction model can be improved by continuously
collecting more sample data of Ldams and the influencing factors of their longevity. At the same time, the research results also emphasize the potential of using 3D laser scanning technology and machine learning algorithm in related engineering fields, including hydrological prediction, geological disaster prediction and structural safety assessment. This study paves the way for further study on the engineering characteristics and failure mechanism of Ldam structure.

6. Conclusion

This paper considers the impact of factors such as Ldam size, material, inducement, and watershed area on Ldam longevity. Multiple intelligent prediction models for Ldam longevity are built using machine learning models such as SVR, MLP, and CatBoost, as well as Stacking ensemble learning technology. A total of 350 sample data sets with relatively accurate and comprehensive consideration of influencing factors were collected from literature. The constructed database was randomly divided into a training data set (80% of the data) and a testing data set (20% of the data). The training data set was used to train the models, and the IGTO optimization algorithm was used to optimize the hyperparameters in the meta-learner of the Stacking prediction framework through multi-objective optimization. This resulted in an intelligent prediction model for Ldam longevity with superior performance. The testing data set was then used to compare the performance of the multiple models and empirical formulas through model evaluation indicators. To further validate the effectiveness and applicability of the constructed model in practical engineering cases, the study collaborated with local enterprise technicians to conduct field investigations and use unmanned aerial vehicle 3D laser scanning to obtain valuable first-hand data on 61 Ldams. Finally, the SHAP method was used for sensitivity analysis to quantify the importance of each feature on Ldam longevity. Based on this, the following conclusions were drawn:

(1) A database of 350 Ldams with longevity data was collected and organized,
and 12 key factors affecting their longevity were selected as the prediction indicator system. The correlation between the various factors was analyzed and the degree of correlation between each factor was quantified, with corresponding explanations given.

(2) An improved optimization algorithm, IGTO, was proposed and its effectiveness was verified through subsequent model performance evaluation tests. It was applied to optimize the hyperparameters of the meta-learner of the Stacking ensemble learning prediction model, and was able to automatically search for the optimal combination of hyperparameters. The results show that IGTO can effectively improve the accuracy of traditional Stacking ensemble learning prediction models, demonstrating the effectiveness and practicality of the IGTO algorithm in the fields of machine learning and intelligent prediction.

(3) The predicted results of the collected Ldam samples show that the intelligent prediction models for Ldam longevity established in this study, based on intelligent optimization algorithms, machine learning, and ensemble learning techniques, have good predictive performance, superior to traditional empirical formulas. It was also found that the current empirical formulas for Ldam longevity prediction have limitations in their applicability.

(4) The prediction results of the IGTO-Stacking model constructed in this research are in good agreement with the actual situation. Among all the models, it has the best predictive performance, with $R^2 = 0.98$, Adj-$R^2 = 0.98$, RMSE=93.98, MAE=48.59, MAPE=0.46, and VAF=98.29. This indicates that the model can better predict the longevity of Ldams, demonstrating the effectiveness and practicality of the method. It provides an effective method for accurate prediction of Ldam longevity and has strong practical significance.

(5) By introducing the SHAP method, sensitivity analysis was conducted for each selected key factor in the research, quantifying the importance of each feature on the longevity of Ldams. The results showed that Ldam size and scale are the most important factors affecting Ldam longevity, especially the width of Ldams, with an importance level of 0.341. Trigger factor and Ldam material have relatively smaller
impacts, totaling only 0.109. This provides an important reference for the design, monitoring, and maintenance of Ldams.

(6) In order to further validate the effectiveness of the constructed model and its applicability in practical engineering cases, this study collaborated with technicians from a local enterprise in Yangquan City, Shanxi Province. Field investigations were conducted, and unmanned aerial vehicle 3D laser scanning technology was used to obtain firsthand data on 61 sets of Ldams. These data include key information such as the actual size, material, inducement, and watershed area of the Ldams. By inputting these data into the constructed intelligent prediction model, we obtained Ldam longevity predictions that are in good agreement with the actual situation, further validating the effectiveness and practicality of the model in future practical engineering cases. In addition, these data not only increase the value of this research but also enrich the currently scarce Ldam database.

Data availability

The data will be obtained from the correspondent on request. The first-hand data of engineering cases have been put in the appendix.

Acknowledgments

This paper gets its funding from Projects supported by National Key Research and Development Program of China - 2023 Key Special Project (No. 2023YFC2907400), National Natural Science Foundation of China(No.52104109), Natural Science Foundation of Hunan Province, China(No.2022JJ40602).

Declaration of Competing Interest

The authors declare no competing interests.
Reference


[34] Abdollahzadeh, B., Soleimanian Gharehchopogh, F., & Mirjalili, S. Artificial gorilla troops


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Engineeringcasedata.xlsx