**Academic Journal of Applied Sciences**
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 2, No. 2, 2026

# A Multi-Source Monitoring Data Fusion and Early Warning Framework Based on LightGBM-SHAP and Weibull Distribution

**Haiben Lin [1], Jiekai Li [1, *], Qiqi Li [2]**

[1] Guangdong Polytechnic Normal University, Guangzhou, China

[2] Shenzhen University, Shenzhen, China

[*] **Corresponding author**: (Email: mugaige@outlook.com)

**Abstract:** This paper addresses the need for multi-source data fusion and risk early warning in slope monitoring by proposing a modeling method that integrates data calibration, feature interpretation, and hierarchical early warning. First, linear regression is used to calibrate fiber-optic displacement data, and data consistency is improved through anomaly removal, multiple interpolation, and wavelet denoising. Subsequently, a LightGBM regression model is constructed to characterize the nonlinear relationship between surface displacement and factors such as rainfall, pore water pressure, microseismic events, and deep-seated displacement. The SHAP method is introduced to explain the contributions of each factor, thereby enhancing the interpretability of the model results. Building on this, Lasso regression is used to screen for key variables and reduce redundant information, followed by the establishment of a probabilistic representation of displacement velocity and graded early warning thresholds based on the Weibull distribution. This method balances prediction accuracy, physical plausibility, and engineering applicability, and can serve as a reference for identifying complex slope deformations, analyzing key controlling factors, and making risk warning decisions.

**Keywords:** LightGBM-SHAP, Graded Early Warning Framework, Lasso Feature Selection.

## 1. Introduction

In complex engineering monitoring scenarios, structural conditions are often influenced by a combination of factors, including environmental disturbances, the evolution of internal damage, and external construction activities. A single monitoring indicator is insufficient to comprehensively reflect the system's dynamic changes, while multi-source data commonly suffers from issues such as noise, missing values, inconsistent units, and complex variable couplings. Consequently, traditional empirical judgments or fixed-threshold methods face certain limitations in identifying anomalies, explaining deformation mechanisms, and supporting early-warning decision-making. Existing research has largely focused on single-model prediction or statistical analysis of indicators, with insufficient integration between data calibration, nonlinear modeling, factor interpretation, and risk classification, making it difficult to establish a comprehensive and generalizable algorithmic framework.

To address this, this paper constructs a comprehensive modeling framework for multi-source monitoring data. The framework first enhances the consistency of multi-source data through data preprocessing and sensor calibration. It then employs the LightGBM model to learn the nonlinear mapping relationship between monitoring variables and target responses, and combines the SHAP method to analyze the contributions of key factors, thereby improving the interpretability of the model results [1]. Building on this foundation, the Lasso method is introduced to screen for core variables and reduce interference from redundant information, while the Weibull distribution is employed to establish graded early warning rules, thereby achieving continuous modeling from data correction and feature identification to risk expression. This paper applies the framework to slope deformation monitoring to validate its applicability in the fusion analysis of multi-source information, including rainfall, pore water pressure, microseismic events, and displacement. It provides a universally applicable algorithmic model approach for engineering disaster risk identification and early warning.

## 2. Data Processing and Sensor Data Calibration Model

### 2.1. Data Source and Preprocessing

The data used in this study come from multiple groups of slope multi-source monitoring systems, including fiber-optic displacement monitoring data and vibrating wire reference displacement data, with environmental disturbance information recorded synchronously, which supports subsequent modeling and analysis.

In the actual monitoring process, due to sensor installation conditions, sampling synchronization errors, and external environmental interference, noise and local anomalies are inevitably present in the raw data. Therefore, before modeling, basic data cleaning is required to ensure temporal consistency across different data sources.

The specific processing steps are as follows:

First, all time series are resampled and the time interval is unified to 10 minutes, so that multi-source data are aligned in the time dimension. Then, the $3\sigma$ criterion is used to identify and remove abnormal jump points to reduce the impact of instantaneous disturbances.

For missing data handling, the MICE multiple imputation method is adopted for estimation. This method performs iterative modeling based on correlations between variables and shows more stable performance compared with single-variable interpolation methods. For event-type variables during non-blasting periods, values are uniformly set to 0 to maintain physical consistency.

Then the displacement sequence is decomposed using a three-level db4 wavelet decomposition, and high-frequency components are suppressed using a soft-thresholding method, so as to reduce noise influence while preserving the overall trend.

After the above processing, the issues of missing values and anomalies are significantly improved in the data, and the dataset meets the requirements of consistency and stability for subsequent modeling.

## 2.2. Linear Regression Sensor Calibration Model

Due to systematic deviation between the fiber-optic displacement sensor and the vibrating wire reference measurement, a linear regression model is used to calibrate the fiber data to reduce the effects of fixed offset and scale error.

Let the original fiber displacement be $x(t)$, and the calibrated displacement be $\hat{y}(t)$, the linear relationship is constructed as follows:

$$\hat{y}(t) = k \cdot x(t) + b \tag{1}$$

Where $k$ represents the scaling coefficient reflecting sensitivity differences, and $b$ represents the zero-offset.

The parameters are estimated using the least squares method, with the optimization objective:

$$\min_{k,b} \sum_{t=1}^{N} \left( y(t) - \hat{y}(t) \right)^2 \tag{2}$$

Where $y(t)$ is the vibrating wire reference displacement data, and $N$ is the total number of samples.

This method has a relatively simple structure, but shows good stability in handling systematic deviations, and can effectively correct long-term drift issues, making different sensor data consistent on the same scale.

## 2.3. Calibration Results and Analysis

The calibration parameters obtained through full time-series fitting are $k = 0.8807,\ b = -0.3697$.

The corresponding linear calibration model is:

$$\hat{y}(t) = 0.8807x(t) - 0.3697 \tag{3}$$

Table 1 shows part of the comparison results before and after calibration.

**Table 1.** Comparison of results before and after calibration

| | | | | | |
|---|---|---|---|---|---|
| Original data x | 7.132 | 18.526 | 84.337 | 123.554 | 167.667 |
| Calibrated data ŷ | 5.911 | 15.946 | 73.907 | 108.446 | 147.296 |

In terms of error metrics, the model performance over the full time series is $R^2 = 0.9992$, RMSE is 2.6823 mm, and MAE is 1.3081 mm.

Figure 1 show the comparison results before and after calibration. It can be seen that the original fiber-optic data has an overall offset compared with the reference data, while after calibration the two sequences are basically consistent in trend. In the sections where displacement changes faster, the deviation reduction is more obvious.

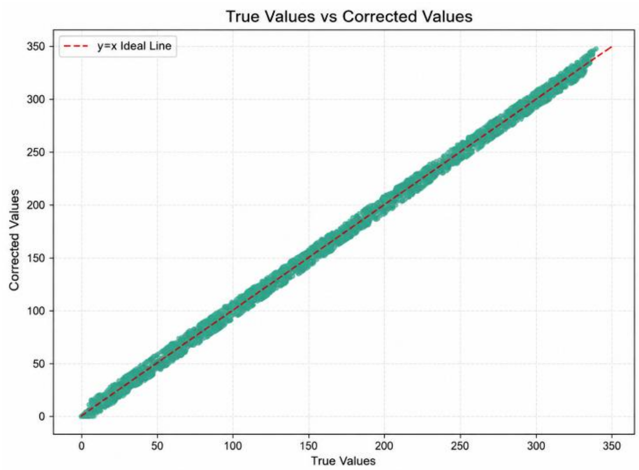


**Figure 1.** Comparison of data after calibration

From the error variation results as shown in Figure 2, the error distribution before calibration is relatively scattered and shows clear bias characteristics. After calibration, the error range is significantly reduced, and the mean absolute error drops by about 94%.

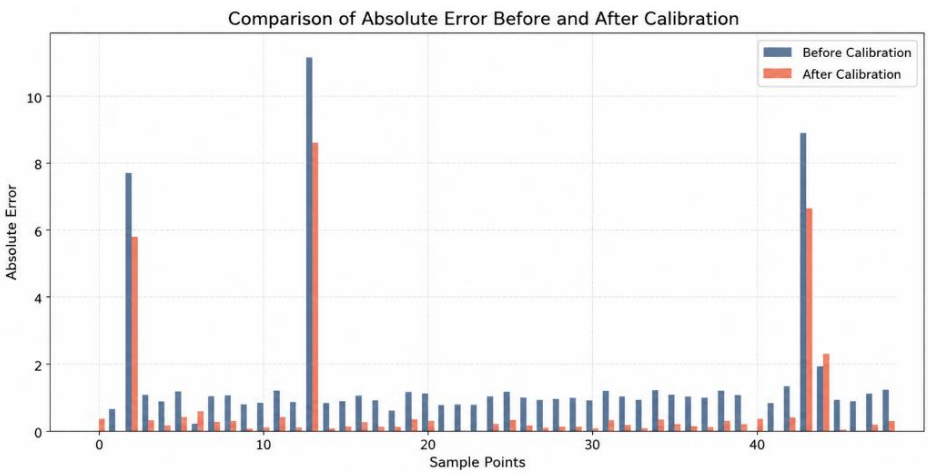


**Figure 2.** Comparison of the error before and after calibration

The results of the residual analysis indicate that most of the residuals are concentrated around zero, with no obvious trend or periodic structure observed, suggesting that the linear correction model has effectively eliminated systematic errors.

# 3. Multi-source Data Modeling and SHAP Interpretability Analysis

## 3.1. Modeling Features and Structural Information Definition

The modeling variables include rainfall, pore water pressure, microseismic event count, deep displacement, and surface displacement, which correspond to hydrological disturbance, internal structural response, and overall deformation output in the slope system.

Rainfall and pore water pressure mainly reflect how external moisture condition changes affect the stability of rock and soil mass. Microseismic events correspond to internal fracture propagation processes. Deep displacement is used to describe accumulated deformation inside the rock mass, while surface displacement represents the overall deformation response.

To reduce the influence of non-stationarity in time series, results from the early change-point detection are introduced as a basis for stage segmentation. The displacement series is divided into three stages: constant velocity, accelerated deformation, and rapid instability. These stages are handled separately in the following analysis to reduce structural mixing in the overall modeling process.

## 3.2. LightGBM Modeling and SHAP Explanation Method

A LightGBM model is used to build a multi-source regression framework to capture the nonlinear mapping relationship between the monitoring variables and surface displacement [2]. This method has good adaptability to coupling relationships among variables and is suitable for complex time-series modeling scenarios.

For model interpretation, the SHAP method is introduced to decompose the prediction results, breaking the model output into additive contributions from each input variable:

$$\hat{S} - S = \sum \phi_j \tag{4}$$

Where $\phi_j$ represents the marginal contribution of the $j$-th feature to the model output, used to measure the influence strength of each factor on displacement variation.

This approach allows local attribution analysis of the black-box prediction results without changing the model structure, making the model output partially interpretable.

## 3.3. Abnormal Structure Analysis and Multivariate Response Characteristics

Based on the isolation forest method, anomaly detection is performed on the multi-dimensional monitoring data to extract abnormal disturbance points in the samples and to observe the synchronous variation relationships among different variables [3].

The results show that the anomaly distributions of different variables are not consistent. Anomalies in pore water pressure and displacement-related variables are relatively concentrated, while rainfall and microseismic events show intermittent fluctuation characteristics. In some time intervals, multiple variables exhibit abnormal responses at the same time, which more likely corresponds to system-level coupled disturbances.

The anomaly detection results indicate significant differences in the distribution of anomalies across different variables. The number of anomalous points is relatively high for variables related to pore water pressure and displacement, while rainfall and microseismic events primarily exhibit intermittent fluctuations. Synchronous anomalies across multiple variables occurred during certain time intervals, reflecting a synergistic response within the slope system. The model training process remained stable, and evaluation metrics showed minimal fluctuation during cross-validation, indicating that the constructed model possesses good stability and generalization capability. Multivariate anomaly analysis suggests that some anomalous events are not generated independently by a single monitoring indicator but are accompanied by synchronous changes in multiple variables, indicating that the slope evolution process is influenced by the combined effects of multiple factors.

## 3.4. Factor Contribution Analysis and Mechanism Interpretation

Based on the LightGBM-SHAP model, the contribution of each input variable is calculated, and the ranking is as follows [4]: deep displacement > pore water pressure > microseismic events > rainfall > blasting disturbance

From a mechanistic perspective, deep displacement mainly reflects internal deformation of the rock-soil mass and is a direct indicator of slope evolution. Changes in pore water pressure correspond to variations in effective stress conditions and have a relatively strong impact on stability. Microseismic events usually occur during crack propagation stages and can provide certain precursor information. Rainfall influences the moisture state of the slope through infiltration processes, but its effect shows a time lag. Blasting disturbance only shows local and short-term influence and contributes less to the overall trend.

This ranking is generally consistent with the hydro–geotechnical coupling process of slopes, indicating that the model results are physically reasonable to some extent.

Figure 3 and 4 show the factor contribution distribution and model prediction results.

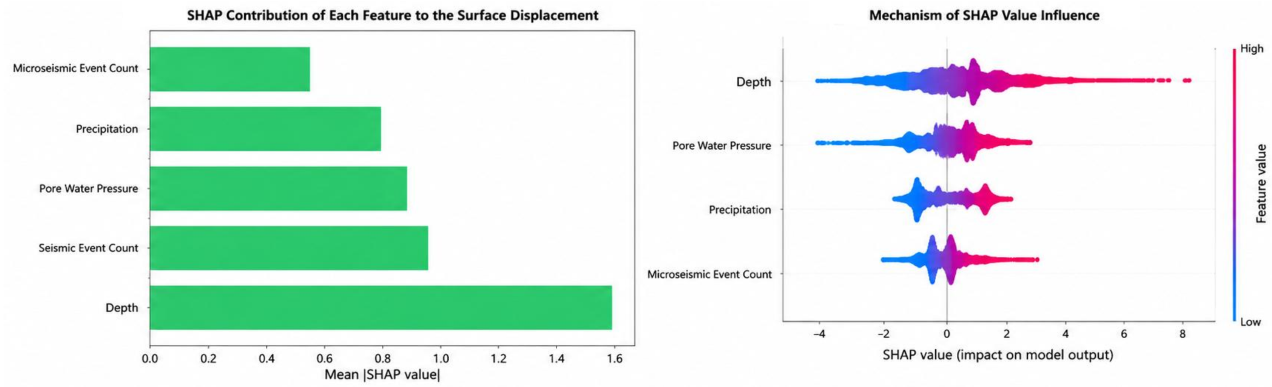


**Figure 3.** Analysis of factor contributions and mechanisms of influence

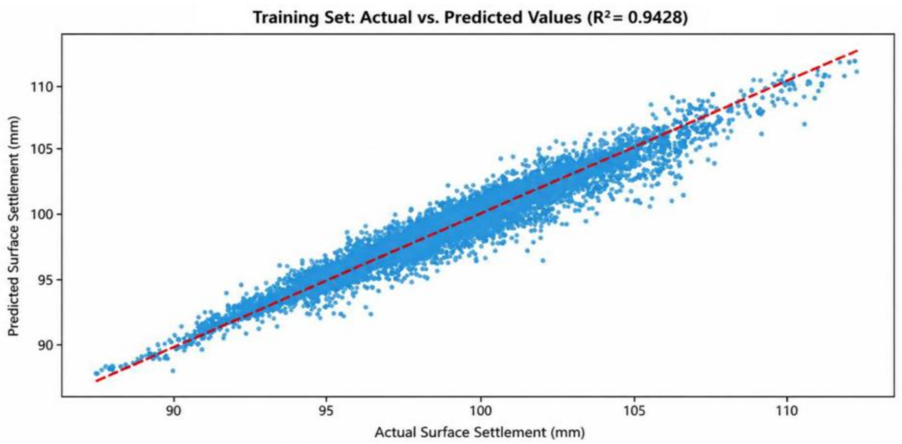


**Figure 4.** Validation of model fit on the training set

### 3.5. Summary

A LightGBM model is constructed based on multi-source monitoring data, and SHAP is used to interpret the model outputs. On this basis, stage-based structural information is introduced to assist modeling, allowing features from different stages to be represented separately.

The results indicate clear differences in the contributions of influencing factors to displacement variation, with deep displacement and pore water pressure playing dominant roles. These findings can provide input support for subsequent risk assessment models.

## 4. Graded Early Warning Model Based on Lasso Selection and Weibull Distribution

### 4.1. Variable Selection and Modeling Idea

Slope displacement is influenced by multiple factors, including rainfall, pore water pressure, microseismic event count, dry–wet infiltration coefficient, and blasting-related variables. However, correlations exist among these variables to different degrees, and directly including all of them in the model may introduce redundant information.

To reduce variable redundancy, Lasso regression is introduced for feature selection. This method applies L1 regularization, which forces part of the coefficients to shrink to zero, achieving sparse variable selection and retaining factors that are more sensitive to displacement variation [5].

Based on this, regression models are constructed using different variable combinations, and $R^2$, RMSE, and MAE are used as comparison metrics to evaluate how changes in feature sets affect prediction performance.

### 4.2. Results of Variable Combination Analysis

Regression models are built based on different feature combinations, and the results are shown in Figure 5.

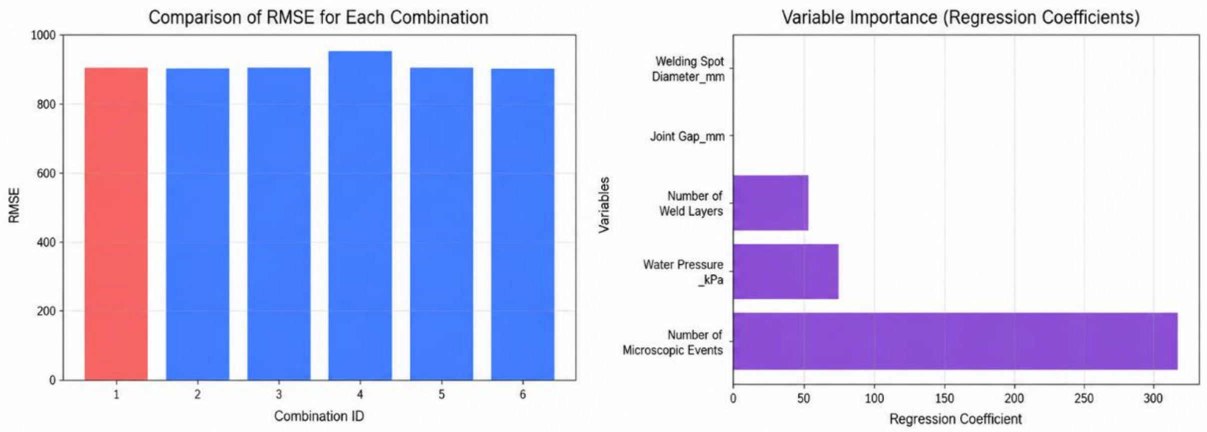


**Figure 5.** RMSE and Variable Importance for Each Combination

The predictive performance varies across different combinations, but the overall trend remains relatively consistent. Combinations including pore water pressure, microseismic event count, and dry–wet infiltration coefficient tend to perform more stably in most cases, while some blasting-related variables do not significantly improve prediction accuracy.

Table 2 presents the quantitative comparison results for each combination.

**Table 2.** Model validation results for different variable combinations

| No. | Variable combination | $R^2$ | RMSE | MAE |
|---|---|---|---|---|
| 1 | Rainfall + Pore water pressure + Microseismic events + Wetting–drying infiltration coefficient + Blasting distance | 0.1417 | 903.8057 | 854.1886 |
| 2 | Rainfall + Pore water pressure + Microseismic events + Wetting–drying infiltration coefficient + Maximum charge per delay | 0.1415 | 903.8998 | 854.7736 |
| 3 | Rainfall + Pore water pressure + Microseismic events + Blasting distance + Maximum charge per delay | 0.1379 | 905.8005 | 852.7454 |
| 4 | Rainfall + Pore water pressure + Wetting–drying infiltration coefficient + Blasting distance + Maximum charge per delay | 0.0419 | 954.9003 | 876.0746 |
| 5 | Rainfall + Microseismic events + Wetting–drying infiltration coefficient + Blasting distance + Maximum charge per delay | 0.1337 | 907.9961 | 868.6210 |
| 6 | Pore water pressure + Microseismic events + Wetting–drying infiltration coefficient + Blasting distance + Maximum charge per delay | 0.1415 | 903.8998 | 854.7736 |

From the results, the optimal combination shows better performance in both $R^2$ and error metrics, indicating that this variable set can better capture the main driving factors of slope displacement changes.

In the Lasso regression results, some variable coefficients approach zero, such as blasting point distance and maximum charge per delay. These variables show weaker influence in the overall modeling process and behave more like short-term disturbance factors.

## 4.3. Weibull Distribution Modeling and Risk Representation

After obtaining the optimal variable combination, a risk modeling indicator is constructed based on staged displacement velocity.

Considering that velocity growth during slope instability shows a clear asymmetric characteristic, the Weibull distribution is used to model the displacement velocity in a probabilistic form. The distribution function is defined as:

$$F(v) = 1 - \exp\left(-\left(\frac{v}{\lambda}\right)^k\right) \tag{5}$$

Where $k$ is the shape parameter and $\lambda$ is the scale parameter.

Based on this distribution, threshold values of velocity under different probability levels are obtained through quantile inversion:

$$v_{th} = \lambda\left(-\ln(1-p)\right)^{1/k} \tag{6}$$

This expression maps observed velocity values into a probability space, providing a basis for subsequent risk classification [6].

## 4.4. Graded Early Warning Results

Based on the 95%, 85%, and 70% quantile levels, a three-level early warning system is constructed, corresponding to yellow, orange, and red warning intervals.

Figure 6 shows the Weibull fitting results and the threshold partitioning.

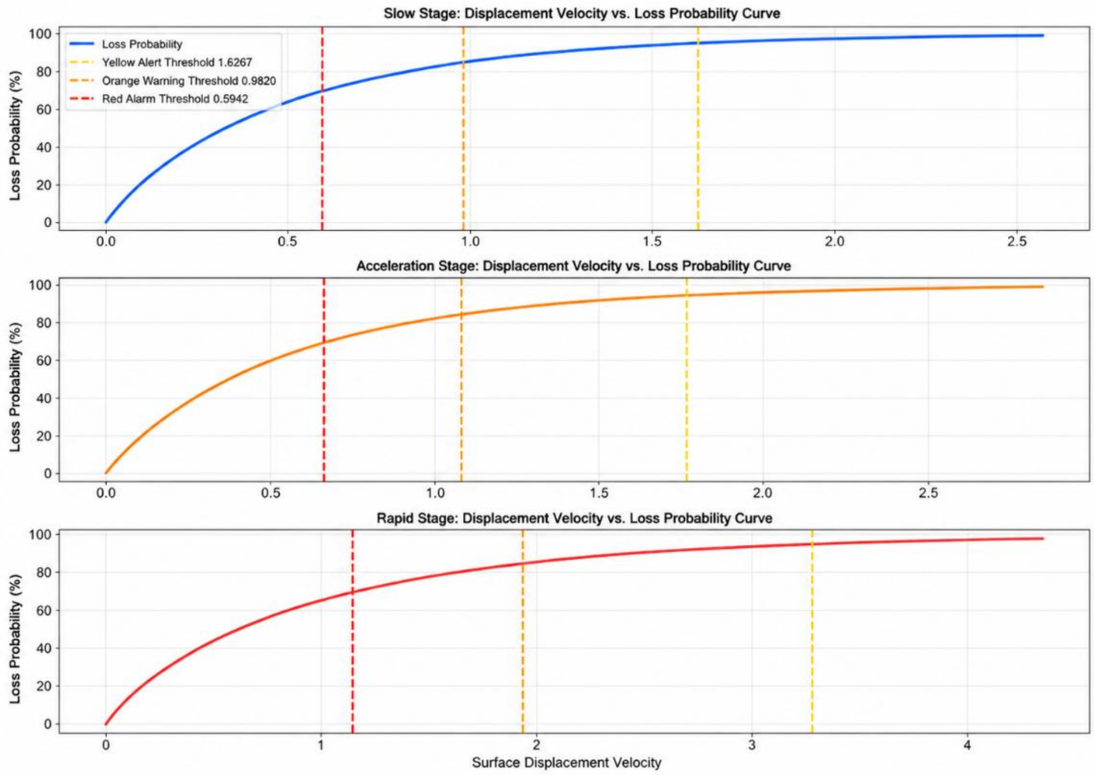


**Figure 6.** Visualization of early warning threshold fitting

From the distribution results, when velocity is low, the probability changes relatively smoothly. When velocity enters a higher range, the cumulative distribution function increases more rapidly, and the risk growth rate also becomes more pronounced.

This behavior is generally consistent with the evolution process of slope stability transitioning into an accelerated failure stage.

Compared with fixed-threshold methods, this approach is based on statistical distribution, so the thresholds are no longer constants but are automatically adjusted according to the data distribution.

## 4.5. Summary

Based on the core variables selected by Lasso, a graded risk representation model is constructed using the Weibull distribution.

The results show that displacement velocity can be well described in a probabilistic form, and different risk levels correspond to different velocity ranges. Compared with fixed-threshold approaches, this method is more flexible in adapting to data fluctuations.

Overall, this model provides a distribution-based approach for slope risk representation and can serve as a reference for subsequent early warning decisions.

# 5. Conclusions

This paper focuses on the application of multi-source monitoring data in engineering risk identification and constructs an algorithmic model framework that integrates calibration, modeling, interpretation, and early warning. The results show that, following linear regression calibration, the fiber-optic displacement data exhibits a high degree of consistency with the reference data, with a model coefficient of determination reaching 0.9992 and the mean absolute error reduced by approximately 94%, effectively enhancing data reliability. LightGBM-SHAP analysis reveals that deep-seated displacement and pore water pressure are the primary factors influencing surface displacement changes, indicating that the model can extract key engineering-relevant information from multivariate coupled relationships. By combining Lasso with the Weibull distribution, displacement velocity can be converted into a graded risk expression, providing a more intuitive basis for early warning decision-making. This framework is not only applicable to slope monitoring but also offers methodological insights for multi-factor risk assessment and dynamic early warning research in the social sciences. Future research could further incorporate data from additional scenarios to enhance the model's generalization capabilities.

# References

[1] Cheng, X. K., & Huang, R. Q. (2025). Explainable flight delay prediction based on LightGBM and SHAP. Computer Times, (11), 26-31. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2025.11.005

[2] Wang, Z. M., Wang, K. F., Li, Y. Z., et al. (2023). A study on forest fire prediction in Yunnan province based on LightGBM and SHAP. Fire Science and Technology, 42(11), 1567-1571.

[3] Zhang, Y., Qin, X., & Peng, H. (2025). Anomaly detection in dam monitoring data based on smoothed thresholds and isolated forests. People's Yellow River, 47(3), 141-145.

[4] Chen, D. L., Sun, D. L., Wen, H. J., et al. (2024). A study on landslide susceptibility using LightGBM-SHAP based on different factor screening methods. Journal of Beijing Normal University (Natural Science Edition), 60(1), 148-158. https://doi.org/10.12205/j.cnki.jbnuns.2024.01.020

[5] Liu, Y. H. (2024). A study on methods for quantifying seismic motion uncertainty in collapse vulnerability analysis based on Lasso regression [Master's thesis]. Institute of Engineering Mechanics, China Earthquake Administration. https://doi.org/10.27490/d.cnki.ggjgy.2024.000088

[6] Wang, Y. N., Yang, J. F., He, Y. S., et al. (2017). Operational risk assessment and early warning of active distribution

networks based on network flow transfer distribution factors. Power System Technology, 41(2), 371-377. https://doi.org/10.13335/j.1000-3673.pst.2016.2815