Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.29992

Prediction of Risk Levels of Embankment Breaches in Small and Medium-Sized Rivers Based on Machine Learning Algorithms

Changtao Deng[1], Xiaowei Li[2], Yiting Wang[3*]

[1]Pearl River Water Resources Research Institute, Guangzhou, China
[2]Hainan Provincial Institute of Modern Emergency Management, Haikou, China
[3]Shenzhen Longhua District Water Pollution Control Center, Shenzhen, China
**Corresponding Author. Email: 624473197@qq.com*

Abstract. The rolling bearings of the train bogie system, as the core basic components of rail transit trains, are widely used in the bogie system and are key components to ensure the safety of train operation and the reliability of train operation. This paper proposes a classification algorithm for rolling bearings in train bogie systems based on KELM-LSTM-Transformer. During the research process, data preprocessing and data analysis are carried out first, followed by correlation analysis, and then this algorithm is compared with multiple machine learning algorithms. From the comparison results, it can be seen that the KELM-LSTM-Transformer significantly outperforms other single models and integrated models in all key evaluation dimensions Its Accuracy is 0.855, Recall is 0.855, Precision is 0.851, and F1 value is 0.847, all of which are the highest among all models. Not only does it far exceed models with relatively weak performance such as the BP neural network with an F1 value of only 0.641 and the GBDT with an F1 value of 0.732, It also outperforms better-performing models such as CatBoost with an Accuracy of 0.811 and an F1 value of 0.792, as well as XGBoost with an Accuracy of 0.792 and an F1 value of 0.795. In terms of the AUC metric for measuring the model's discrimination ability, the algorithm's value of 0.983 is also ahead of the integrated models such as CatBoost with an AUC of 0.974 and XGBoost with an AUC of 0.97, demonstrating a stronger ability to distinguish positive and negative samples. The proposal of this algorithm provides an effective technical solution for improving the accuracy of condition monitoring and fault classification of rolling bearings in the bogies of rail transit trains, which is of great practical significance for ensuring the safe and stable operation of high-speed trains and promoting the intelligent upgrade of the rail transit field.

Keywords: Train bogie system, rolling bearing, LSTM, Transformer

1. Introduction

Small and medium-sized rivers, as weak links in the flood control system of river basins, often suffer from severe loss of life and property as well as ecological damage due to sudden embankment breaches [1]. Due to the impact of climate change, the frequency of extreme rainfall events has

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.29992

increased. Coupled with the low construction standards and insufficient maintenance of some small and medium-sized river embankments, the prevention and control of embankment breach risks are facing severe challenges. Traditional risk assessment of embankment breaches mostly relies on empirical formulas and static monitoring data, making it difficult to capture the dynamic changes of hydrological processes and the coupling effects of multiple factors. The timeliness and accuracy of early warning are limited. Constructing a model that can integrate multiple influencing factors and dynamically predict the risk of embankment failure is of great practical significance for enhancing the disaster prevention and control capacity of small and medium-sized rivers [2].

Machine learning algorithms provide a new technical approach for predicting the risk of river embankment breaches [3]. Compared with traditional methods, machine learning models can automatically mine complex correlations among multi-dimensional data such as hydrology, geology, and engineering, and achieve rapid assessment of risk levels through historical data training [4]. For instance, the decision tree model can clearly present the mechanism of action of key influencing factors, and neural networks can fit nonlinear mapping relationships, effectively enhancing the prediction accuracy. The application of these algorithms not only reduces the reliance on human experience but also integrates real-time monitoring data to achieve dynamic early warning, providing scientific support for flood control decision-making and demonstrating significant advantages in river basin disaster risk management.

Although existing machine learning models have achieved certain results in the prediction of breach risks, most of them are difficult to fully capture the long-term dependencies of time series data and the interaction of spatial features [5]. To this end, this paper proposes the Transformer-BigRU classification algorithm, which combines the self-attention mechanism of Transformer with the bidirectional temporal modeling ability of BiGRU to enhance the ability to capture the dynamic evolution process of multiple factors. This algorithm can not only identify key influencing factors and their time-varying weights through the self-attention mechanism, but also extract the temporal characteristics of hydrological processes with the help of the bidirectional gated cyclic unit, thereby more accurately predicting the risk level of embankment failure in small and medium-sized rivers and providing technical support for the construction of an efficient disaster early warning system.

2. Data sources

This dataset contains 726 records of risk assessment for embankment breaches in small and medium-sized rivers, including 10 characteristic variables and 1 predictor variable. The characteristic variables include the cumulative rainfall in the past 72 hours, the current water level of the river, the ratio of the warning water level of the river, the integrity score of the embankment structure, the age of the embankment, the vegetation coverage rate of the embankment, the average temperature in the past 30 days, the flow velocity of the river, the soil moisture content of the basin, and the maintenance frequency of this section of the river in the past 5 years. The predictor variable is the risk level of embankment failure (0-4 levels, with higher levels indicating higher risks). The data collection time span is nearly two years. This data can objectively reflect the potential risk level of sudden breaches in small and medium-sized rivers, providing data support for the construction of related machine learning models. Some datasets are selected for display, as shown in Table 1.

DOI: 10.54254/2755-2721/2025.29992

Table 1. A partial dataset

Cumulative rainfall	Current water level	Warning water level ratio	Structural integrity	Age	Vegetation coverage rate	Average temperature	River flow rate	Soil moisture content	Maintenance frequency	Risk level of embankment breach
177	2.84	0.83	61.4	16.9	50.5	19.5	1.01	27	7	2
27.5	2.93	1	72	13.2	46.9	26.7	1.57	92.5	3	2
31	5.37	0.93	72	25	14.4	24.4	1.17	21.8	4	2
36.5	5.96	0.54	83.4	40.2	33.1	17.8	1.29	56.8	7	1

3. Method

3.1. Transformer

The Transformer algorithm is based on the deep learning model proposed in 2017, which is based on the self-attention mechanism. It completely breaks away from the sequence-dependent limitations of recurrent neural networks and significantly improves the training efficiency through parallel computing [6]. The core lies in the self-attention mechanism, which can dynamically capture global dependencies by calculating the correlation weights among the elements in the sequence. The model consists of two parts: the encoder and the decoder. The encoder converts the input sequence into a feature vector containing global context information, while the decoder generates the target result based on the encoder output and its own sequence [7]. When dealing with time series data, the Transformer injects sequence order information through position encoding, enabling the model to perceive the temporal position relationship of elements. The network structure of the Transformer is shown in Figure 1.

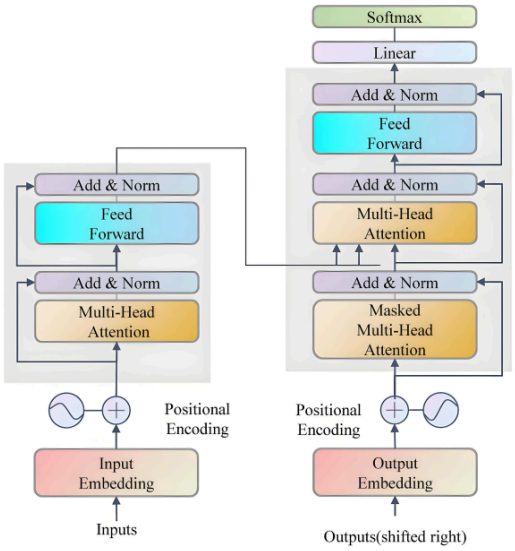

Figure 1. The network structure of the Transformer

DOI: 10.54254/2755-2721/2025.29992

3.2. BiGRU

The BiGRU algorithm is the abbreviation of Bidirectional Gated Recurrent Unit, which is an improved version of recurrent neural networks. It adds a reverse computing mechanism on the basis of traditional GRU. GRU controls the flow and forgetting of information by updating gates and resetting gates, solving the gradient vanishing problem of traditional RNNS and effectively capturing the long-term dependencies of time series data [8].

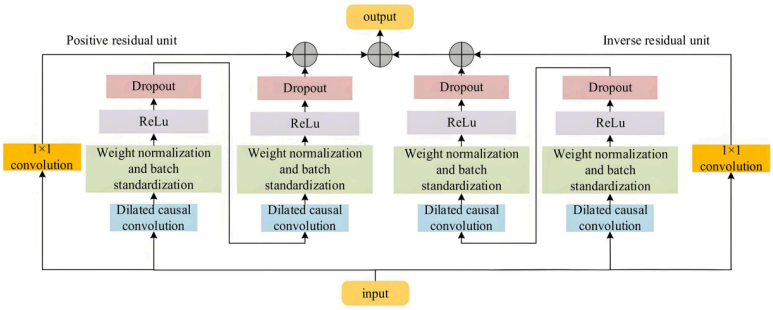

Figure 2. The network structure of BiGRU

BiGRU consists of a forward GRU and a reverse GRU. The forward GRU processes sequence information in chronological order, while the reverse GRU processes it in reverse order. The outputs of both are concatenated or fused to obtain feature representations that contain past and future contexts. This bidirectional structure enables it to focus on both historical evolution patterns and future development trends when processing time series data, significantly enhancing the sensitivity to key time nodes in the sequence [9]. The network structure of BiGRU is shown in Figure 2.

3.3. Transformer-BiGRU

The Transformer-BiGRU classification algorithm is a hybrid architecture that integrates the advantages of both models and is specifically designed for handling time series classification tasks influenced by multiple factors [10]. This algorithm first utilizes the bidirectional cyclic structure of BiGRU to extract the local temporal features of the input sequence and capture the short-term dependencies of dynamic changes such as hydrological processes. Then, the temporal features output by BiGRU are input into the Transformer encoder. Through the self-attention mechanism, the correlations among global features are modeled to identify the key influencing factors and their time-varying weights. In the feature fusion stage, the model adopts a skip connection strategy to retain feature information at different levels and enhance the model's learning ability for complex patterns. For the task of predicting the risk of river breaches, this algorithm can not only capture the dynamic change trends of indicators such as water level and rainfall through BiGRU, but also mine the coupling effects among multiple factors with the help of Transformer. Finally, it outputs the risk level classification results through the fully connected layer to achieve precise assessment of the risk of breaches.

4. Result

When dividing the dataset, the proportion of the training set is 0.7. The input feature dimension is determined by the data, and the number of categories is determined by the number of unique labels in the data. In terms of the model, the number of input channels is consistent with the feature

DOI: 10.54254/2755-2721/2025.29992

dimension. The maximum position encoding is set to 512, the number of self-attention mechanism heads is 4, the number of key channels for each head is 32, and the total number of key channels is 128. The GRU layers are respectively set to 6 and 10, and the Dropout layer ratio is 0.01. In terms of training parameters, the optimizer selects Adam, with a maximum of 200 training rounds, a batch size of 256, an initial learning rate of 0.01, a learning rate decline factor of 0.1, a decline period of 80, an L2 regularization coefficient of 0.001, a gradient clipping threshold of 10, and an automatic selection of the execution environment.

The results of the output comparative test are shown in Table 2 and Figure 4.

Table 2. The results of the comparative experiment

Model	Accuracy	Recall	Precision	F1	AUC
Decision tree	0.775	0.775	0.786	0.779	0.829
Naive Bayes	0.862	0.862	0.85	0.85	0.97
XGBoost	0.849	0.849	0.839	0.843	0.947
KNN	0.642	0.642	0.597	0.615	0.843
CatBoost	0.876	0.876	0.84	0.857	0.956
AdaBoost	0.771	0.771	0.801	0.771	0.899
Transformer-BiGRU	0.908	0.908	0.907	0.905	0.979

From the results of the model performance evaluation, the Transformer-BiGRU model proposed in this paper demonstrates significant advantages in all key indicators, and its overall performance is comprehensively superior to traditional and classic machine learning models such as decision trees, Naive Bayes, XGBoost, KNN, CatBoost, and AdaBoost. In terms of core classification performance metrics, the Accuracy, Recall, Precision and F1 values of Transformer-BiGRU are all the highest among all models. Not only does it significantly outperform the underperforming KNN model - all its metrics are below 0.65. It also significantly outperforms decision trees and AdaBoost - their accuracies are 0.775 and 0.771 respectively, and their F1 values are 0.779 and 0.771 respectively. At the same time, it exceeds the better-performing CatBoost and XGBoost - the former has an Accuracy of 0.876 and an F1 value of 0.857. The latter has an Accuracy of 0.849 and an F1 value of 0.843, and is only slightly higher than Naive Bayes' 0.85 in Precision. In terms of the AUC metric for evaluating the discriminative ability of models, although Transformer-BiGRU is relatively close to Naive Bayes, it still leads by a narrow margin and is significantly higher than XGBoost, CatBoost and other models. This fully demonstrates that the model has stronger capabilities in feature extraction, category differentiation and prediction stability, effectively overcoming the limitations of traditional models in capturing complex data patterns.

The confusion matrix of Transformer-BiGRU is shown in Figure 4.

DOI: 10.54254/2755-2721/2025.29992

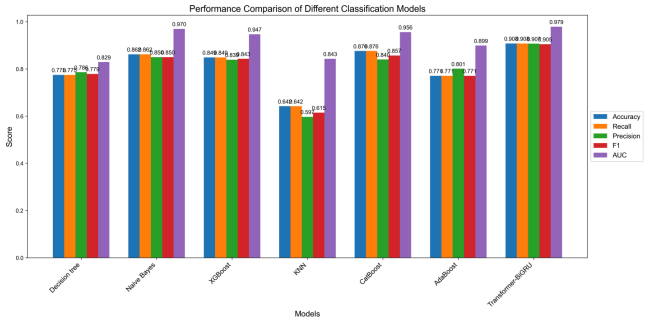

Figure 3. The comparison bar chart of the indicators of each model

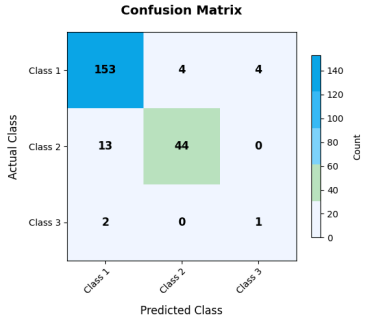

Figure 4. The confusion matrix of Transformer-BiGRU

5. Conclusion

Small and medium-sized rivers are the weak links in the flood control system of river basins. When their banks break suddenly, it often causes significant casualties, property losses and ecological damage. To this end, this paper constructs a Transformer-BigRU classification algorithm. This algorithm integrates the self-attention mechanism of Transformer and the bidirectional time series modeling characteristics of BiGRU, which can capture the multi-factor dynamic evolution process more efficiently. Through conducting comparative experiments with various machine learning algorithms, it was found that the Transformer-BiGRU model performed outstandingly in all key performance indicators. Overall performance far exceeds that of traditional and classic machine learning models such as decision trees, Naive Bayes, XGBoost, KNN, CatBoost, and AdaBoost. In terms of core classification indicators, its Accuracy, Recall, Precision and F1 value all rank first: Not only does it significantly outperform the KNN model with Accuracy and Recall both below 0.65, but it is also notably superior to the decision tree and AdaBoost models with accuracies of 0.775 and 0.771 respectively and F1 values of 0.779 and 0.771 respectively. It simultaneously exceeds the CatBoost model with an Accuracy of 0.876 and an F1 value of 0.857, as well as the XGBoost model with an Accuracy of 0.849 and an F1 value of 0.843. Only slightly higher than Naive Bayes' 0.85 in Precision; In terms of the AUC index reflecting the discriminative ability of the model, although this model is similar to Naive Bayes, it still leads by a narrow margin and is significantly higher than other models. This indicates that Transformer-BiGRU has more advantages in feature extraction, category differentiation and prediction stability. It can effectively make up for the deficiencies of traditional models in capturing complex data patterns, providing reliable technical support for the precise early warning of embankment breaches in small and medium-sized rivers, and has important practical significance for improving the flood control safety guarantee capacity of river basins.

DOI: 10.54254/2755-2721/2025.29992

References

[1] Yukawa, Chihiro, et al. "An Intelligent Water Level Estimation System Considering Water Level Device Gauge Image Recognition and Wireless Sensor Networks." Journal of Sensor and Actuator Networks 14.1 (2025): 13.
[2] Ito, Takehiko, et al. "Nowcast and forecast modeling of water-level profiles in rivers based on improved DIEX-flood and deep learning." Journal of JSCE 10.1 (2022): 307-314.
[3] Liu, Xu, et al. "Assessment of typhoon disaster loss based on the factor analysis-random forest model." Journal of Physics: Conference Series. Vol. 2718. No. 1. IOP Publishing, 2024.
[4] Kpiebaya, Prosper, Abdul-Ganiyu Shaibu, and Eliasu Salifu. "Machine learning-based modeling of suitable dam sites in Northern Ghana." H2Open Journal 8.4 (2025): 271-290.
[5] Huang, Y., H. H. Ba, and S. Yu. "Hydroinformatics in China: overall developments and showcase of accomplishments in the Changjiang River basin." Michael Abbott's Hydroinformatics: Poiesis of New Relationships with Water. IWA Publishing, 2022. 75-92.
[6] Xia, Jun, and Jin Chen. "A new era of flood control strategies from the perspective of managing the 2020 Yangtze River flood." Science China Earth Sciences 64.1 (2021): 1-9.
[7] Werritty, Alan, et al. "Use of multi-proxy flood records to improve estimates of flood risk: Lower River Tay, Scotland." Catena 66.1-2 (2006): 107-119.
[8] Wei Wenbo, et al. "Research on Fault Diagnosis of Train Bearings Based on Spectral Recursive Image Coding Technology." Rolling Stock (1002-7602) 63.4 (2025).
[9] Yukawa, Chihiro, et al. "An Intelligent Water Level Estimation System Considering Water Level Device Gauge Image Recognition and Wireless Sensor Networks." Journal of Sensor and Actuator Networks 14.1 (2025): 13.
[10] Masumoto, Takao, Takeo Yoshida, and Tomijiro Kubota. "An index for evaluating the flood-prevention function of paddies." Paddy and Water Environment 4.4 (2006): 205-210.