Prediction of Dam Seepage through a Machine Learning Technique and Its Application to Dam Diagnosis

Conference Paper

Author(s):
Okabe, Hokuto; Oshita, T.; Suzuki, M.; Inoue, K.

Publication date:
2024

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000676002

10th International Symposium on Hydraulic Structures Zurich, Switzerland 17-19 June 2024
ISSN 0374-0056 DOI: 10.3929/ethz-b-000676002

Prediction of Dam Seepage Through a Machine Learning Technique and Its Application to Dam Diagnosis

H. Okabe[1], T. Oshita[2], M. Suzuki[1] & K. Inoue[1]
[1]*Graduate School of Agricultural Science, Kobe University, Kobe, Japan*
[2] *Hyogo Prefectural Government, Kobe, Japan*
E-mail: mornel@kobe-u.ac.jp

Abstract: *This study proposes a machine learning model for dam management to achieve low-cost and rapid dam diagnosis based on observation data related to the seepage rate. This study employs extremely randomized trees (ERT), which is one of the machine learning techniques derived from the decision tree, as a learning algorithm. This relatively new algorithm is anticipated to yield an optimal outcome using a variety of observation information compared to other widely used tree-based approaches, such as the decision tree and random forest. In this study, transitionally observed seepage rates at actual rock-fill dams located in the Tohoku province in Japan were used to build the training and testing data. In addition to the measured seepage rate, 14 types of rainfall data including the rainfall from one day through 14 days before, the water level in the reservoir, and the temporal gradient of water level, were employed. This is because the seepage rate may be influenced not only by the rainfall event of the day but also by the status of the water level in the reservoir. The elapsed time since the floods test was also treated as label data in the learning process. These 44 variables were the data set in the learning model and were treated as features for seepage prediction. As a result, it was revealed that the water level in the reservoir, the rainfall on the day, and the temporal gradient of the water level are important for achieving highly accurate prediction.*

Keywords: *Dam Seepage, Features, Machine Learning, Soundness Assessment*

1. Introduction

From the inception of agriculture to today, the symbiotic relationship between human life and agriculture has been integral. However, the demand for water, particularly for irrigation and food production, poses a significant challenge to freshwater resources, with agriculture accounting for approximately 70% of global freshwater withdrawals (UNESCO World Water Programme 2012). In managing water resources within these limitations, dams play a crucial role in civilization by supplying water periodically from reservoirs based on demand patterns. Upon construction, dams and their components undergo aging at varying rates. While some dams may remain structurally sound for millennia, others may develop cracks and leakage in less than a decade. Recognizing these critical points, the International Commission on Large Dams (ICOLD 2017) emphasizes a shift towards prioritizing the restoration of structural and operational safety in existing dams over designing and constructing new ones.

With approximately 2,700 dams in Japan, the Ministry of Agriculture, Forestry and Fisheries directly oversees 189 dams, 45 of which have surpassed 50 years since construction. The aging infrastructure raises concerns about functional deterioration, necessitating thorough dam health assessments. Additionally, the increasing frequency of heavy rainfall disasters due to global climate change and concerns about seismic resilience underscore the urgency of reassessing dam safety. Monitoring dam health involves focusing on different areas of the dam body, relevant components, and physical processes such as structural deformations, water infiltration, and corrosion. Basic monitoring data include reservoir water levels, rock displacement, phreatic surfaces, and surface deformations, critical for flood prevention and slope stability (Calamak and Yanmaz 2014, Siacara et al. 2020). Seepage discharge flow from filter drains is also crucial information, particularly for earth-fill or rock-fill dams, serving as a vital indicator of dam health (ICOLD 2001) since the long-term changes in seepage discharge rate signal alternation in the original functionality of the dam.

Considering the challenges in constant dam health assessments and the impracticality of renovating all dams due to management costs, this study explores the potential of using soft computing approaches, specifically machine learning, to predict seepage discharge rates. Soft computing approaches have demonstrated success in solving real-world problems and identifying optimal values across various scientific disciplines (Masmoudi et al. 2020). Notably, artificial intelligence (AI) and machine learning (ML) have been effectively applied in order to assess the status of the dam using a wide variety of methodologies (e.g., Wang et al. 2009, Hipni et al. 2013, Roushangar and Alipour 2018, Belmokre et al. 2019, Li et al. 2021, Sorkhabi et al. 2022).

The objective of this study is to develop a ML model for predicting the seepage rate of a dam, leveraging monitoring data to assess the health of the dam. By selecting appropriate features for seepage estimation, we aim to propose a ML model that establishes a benchmark for evaluating the dam's health efficiently and cost-effectively through the comparison of estimated and observed values.

2. Materials and Methods

2.1. Extremely Randomized Trees

Extremely randomized trees (ERT) (Geurts and Wehenkel 2006) is a ML method that determines values through multiple independently trained decision trees, introducing randomness into the learning process. A decision tree, as illustrated in Figure 1, is a model that branches data based on the magnitude of the branching criteria S_t for the feature X_j. When a node t branches into t_L and t_R, the branching criteria are selected to minimize the weighted variance of the data, following and ensuring:

$$minimize\ \Delta I(t) \tag{1}$$

$$\Delta I(t) = \frac{N(t_R)}{N(t)} I(t_R) + \frac{N(t_L)}{N(t)} I(t_L) \tag{2}$$

$$I(t) = \sum_{y_i \in D_t} \frac{(y_i - \bar{y})^2}{N(t)} \tag{3}$$

Here, $N(t)$ represents the number of data points at node t, D_t represents the data at node t, y_i is the target variable, and $\bar{y}$ denotes the mean of y_i. ERT selects candidate values for branching criteria randomly from each feature and repeats branching up to the maximum depth, outputting the mean of the terminal data as the final prediction.

The estimation result f of ERT is determined by:

$$f = \frac{1}{M} \sum_{m=1}^{M} h_m \tag{4}$$

where h_m represents the output values for each decision tree, and M denotes the number of trees, set to 1000 in this study. ERT allows the calculation of feature importance for each feature. The importance $Imp(X_j)$ of features X_j is represented by:

$$Imp(X_j) = \frac{imp(X_j)}{\sum_{j=1}^{J} imp(X_j)} \tag{5}$$

$$imp(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in \phi_m} \frac{N(t)}{N} \Delta I(t,j) \tag{6}$$

Here, J is the number of features, M is the number of trees, ϕ_m represents all branching points in decision tree m, N is the total data count, and $\Delta I(t, j)$ indicates the $\Delta I(t)$ at the time of branching using X_j (Louppe 2015). Due to its relatively high accuracy (Inoue and Suzuki 2020) and the ease of understanding the concept of learning through decision trees, this study employed ERT as a ML algorithm to estimate seepage in dam seepage rate prediction.

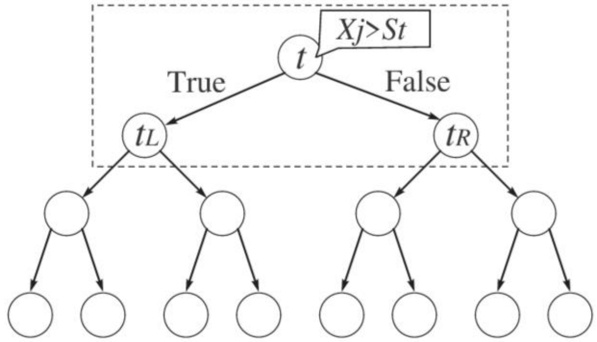

Figure 1. Image of decision tree.

2.2. Data Sets

This study focuses on Dam A: a central core rock-fill dam located in the Tohoku region of Japan. Dam A was constructed from 1975 to 1988 and experienced the impact of the Tohoku-Pacific Ocean Earthquake on March 11, 2011. Subsequently, repair work around the dam crest was carried out in 2014 and 2015, a water level rise test was conducted in 2016, and the dam resumed operation in 2017. The data utilized in this study spans from July 1, 1986, to March 23, 2021, recorded at one-day intervals, including daily rainfall, reservoir water levels, and seepage rate data, as shown in Figure 2. The upper part of the figure represents the features used for estimating seepage, which includes daily rainfall and reservoir water levels, while the lower part illustrates the trend of seepage rate. Due to instrument failures caused by the earthquake, measurements resumed in July 2012. Considering both the instrument failures and the earthquake's impact on seepage rate, the data is segmented into pre-earthquake and post-earthquake periods for separate evaluations.

Figure 2 reveals not only abrupt changes in seepage rate due to rainfall but also that more than one-third of the days after the earthquake show no significant variation in seepage rate. By capturing the trend of seepage rate variations, it is believed that simplifying the estimation model could lead to improved accuracy. Therefore, in this study, two types of target variables were evaluated: (1) the seepage rate and (2) the difference value between the seepage rate and the previous day's seepage rate, referred to as 'the seepage change'. Seepage estimation in pre-earthquake data is denoted as Case X, the estimation of seepage change in pre-earthquake data as Case X', seepage estimation in post-earthquake data as Case Y, and the estimation of seepage change in post-earthquake data as Case Y'.

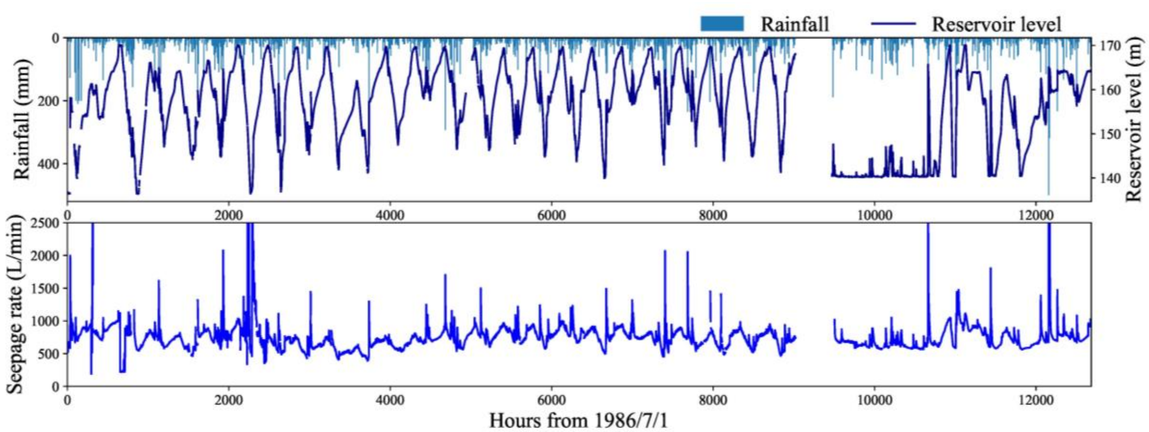

Figure 2. Input data at Dam A.

3. Results and Discussion

3.1. Feature Importance

In the process of feature selection, the evaluated features by determining the importance of each feature has been narrowed down. In this study, three types of features involving rainfall, reservoir water levels, and reservoir water level gradients were considered. The features for evaluation are presented in Table 1. Here, r_n represents the cumulative sum of daily rainfall from the target day to n days ago, l_n is the reservoir water level from the target day to n days ago, gl_n is the difference between the target-day reservoir water level and the reservoir water level n days ago, and *time* indicates the elapsed days from the start of measurement to the target day. Using data from the target day to 14 days ago, a total of 44 features were employed. These features were learned through ML for their relationships associated with seepage rate and seepage change, and the importance of each feature in the regression model was determined.

Figure 3 illustrates the total importance of features for each Case. The importance of *time* decreased after the earthquake compared to before, likely due to the smaller number of data points after the earthquake, resulting in smaller long-term changes. Additionally, in seepage change estimation, the importance of gl_n increased compared to seepage estimation, while the importance of *time* decreased. Since the seepage rate is significantly influenced by the target-day reservoir water level, estimating seepage change, which examines the change in seepage, is considered to be greatly affected by changes in reservoir water levels. Furthermore, as seepage change is determined by the difference from the previous day's seepage rate, long-term effects are presumed to be reduced.

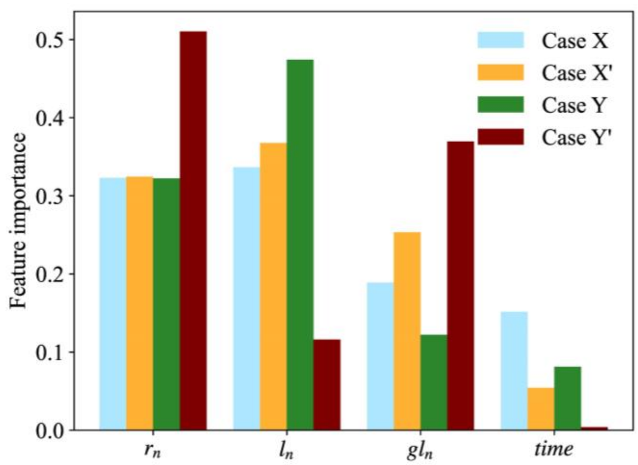

Figure 3. Total importance of features.

Figure 4 shows the importance of features before the earthquake, while Figure 5 displays the importance of features after the earthquake. Smaller values of n, indicating features closer to the target day, exhibited higher importance. Specifically, in seepage estimation, the reservoir water level on the target day showed high importance, while in seepage change estimation, the importance of the reservoir water level gradient from the previous day was particularly high. Therefore, the closer the day is to the estimation target, the greater the impact on the seepage rate, highlighting the significant influence of reservoir water levels.

In cases other than Case Y, smaller values of n resulted in higher importance for r_n, with r_2 showing the highest value. Therefore, it was indicated that a 2-day rainfall period significantly influenced the seepage rate in Dam A. On the contrary, in Case Y, larger values of n led to higher importance for r_n, with r_5 showing the highest value. The observed increase in the seepage rate due to rainfall tended to decrease more gradually than during the ascent, suggesting that rainfall has a long-term impact on the seepage rate. In the post-earthquake period, the shorter observation period resulted in fewer cases of observed rainfall. Consequently, an excessive learning of the long-term effects of rainfall led to an increased importance of longer-term rainfall after the earthquake in Case Y. Moreover, since the impact of long-term rainfall on seepage change was smaller compared to the seepage rate, the importance of a 2-day rainfall period with a significant influence was suggested in Case Y'.

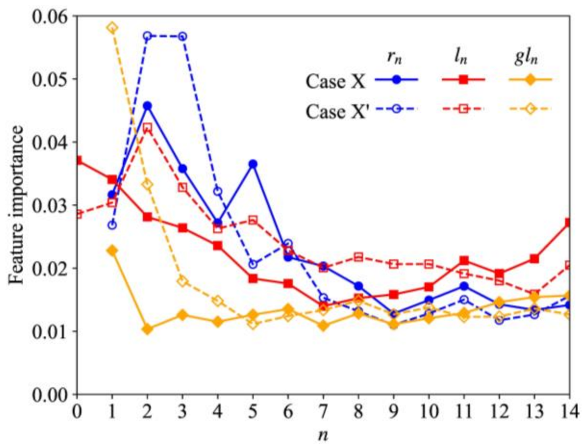

Figure 4. Importance of features before the earthquake.

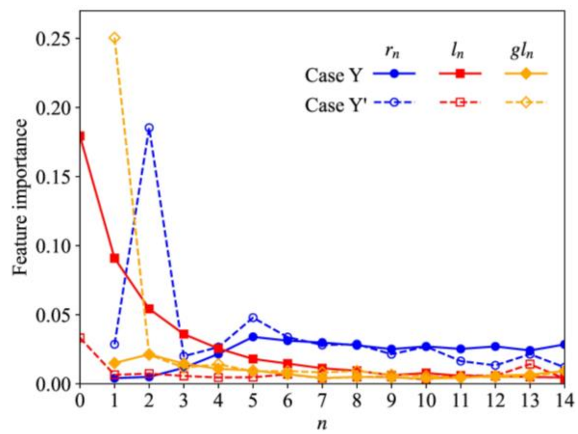

Figure 5. Importance of features after the earthquake.

3.2. Selection Method

ML is a methodology that involves constructing a model using training data and evaluating its performance on test data, without using the same data for training. Considering practical management, the dataset, excluding the most recent 365 days, has been partitioned into training and test data using cross-validation (Chen 2020) after removing missing data. Cross-validation involves randomly shuffling the data into k subsets, as illustrated in Figure 6. One subset is then used as test data, and the rest serves as the training data. The model is evaluated based on the estimation accuracy in k different test data. In this study, the number of splits k for cross-validation was set to 10, and the estimation accuracy on test data was evaluated using the Root Mean Squared Error (*RMSE*). *RMSE* is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

where n represents the number of test data, y_i represents the observed value, and $\hat{y}_i$ represents the estimated value.

Features were selected from those listed in Table 1, up to 14 days before the target day. For each feature type, features in ascending order of n (closer to the target day) were incrementally added and *RMSE* were minimized to determine the optimal features. Following the order of importance demonstrated in Figure 3, the features, namely precipitation r_n, reservoir water level l_n, and reservoir water level gradient gl_n, were evaluated in that sequence. The optimal features for each case were determined for each type of feature.

Table 1. The number and types of features to be evaluated.

Type	Representation of features	Number of features
Rainfall	r_1, r_2, … r_{13}, r_{14}	14
Reservoir water levels	l_0, l_1, … l_{13}, l_{14}	15
Reservoir water level gradients	gl_1, gl_2, … gl_{13}, gl_{14},	14
Elapsed days	*time*	1

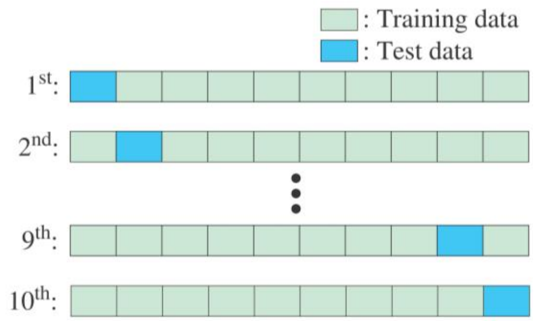

Figure 6. Image of features cross-validation.

3.3. Feature Importance Selection

First, rainfall r_n in ascending order of n were incrementally added and *RMSE* were minimized to determine the optimal number of features. In this case, reservoir water levels l_n and reservoir water level gradients gl_n utilized features up to 14 days before the target day, as shown in Table 2. Figure 7 illustrates the results of *RMSE* concerning the total count of rainfall r_n. In the case of estimating the seepage change, errors significantly decreased with a 2-day rainfall period, and rainfall beyond 2 days introduced noise in the estimation, exhibiting a similar trend before and after the earthquake. On the other hand, in the case of estimating the seepage rate, an improvement in accuracy due to long-term rainfall was observed compared to estimating the seepage change.

Table 2. Features used in rainfall r_n evaluation.

	Case X	Case X'	Case Y	Case Y'
l_n	$l_0 \sim l_{14}$	$l_0 \sim l_{14}$	$l_0 \sim l_{14}$	$l_0 \sim l_{14}$
gl_n	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$

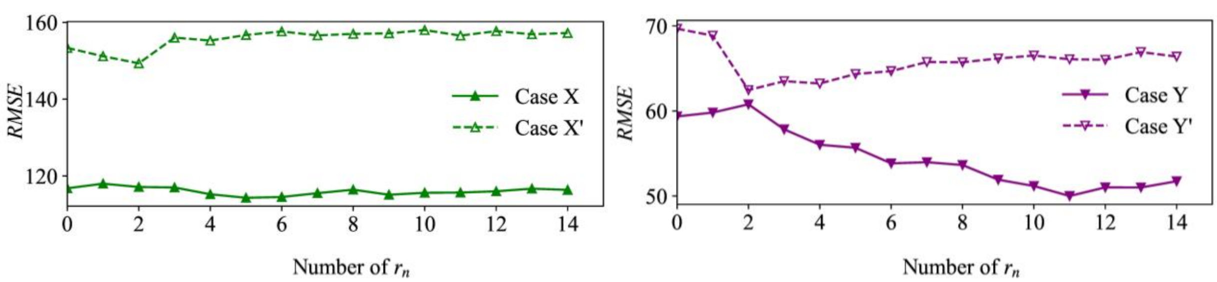

Figure 7. Estimation accuracy concerning the total count of rainfall r_n.

Next, reservoir water levels l_n in ascending order of n were incrementally added and *RMSE* were minimized to determine the optimal number of features. In this case, rainfall r_n and reservoir water level gradients gl_n utilized the optimal features obtained from Figure 7, as well as features up to 14 days before the target day, as shown in Table 3. Figure 8 presents the results of *RMSE* concerning the total count of reservoir water levels l_n. In seepage estimation, errors were significantly reduced with the target-day reservoir water level, highlighting its importance. Additionally, Case X showed a gradual reduction in errors with the inclusion of long-term reservoir water levels. From Figure 2, periodic variations in reservoir water levels were observed before the earthquake, and learning the periodicity of reservoir water levels contributed to the improvement in the accuracy of estimating seepage.

Table 3. Features used in reservoir water levels l_n evaluation.

	Case X	Case X'	Case Y	Case Y'
r_n	$r_1 \sim r_5$	$r_1 \sim r_2$	$r_1 \sim r_{11}$	$r_1 \sim r_2$
gl_n	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$	$gl_1 \sim gl_{14}$

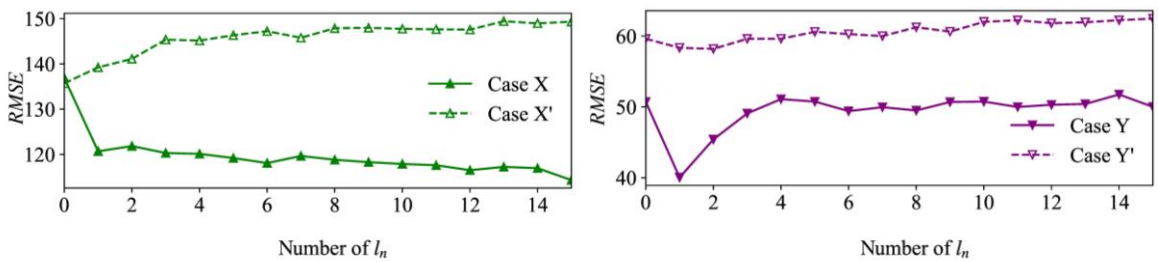

Figure 8. Estimation accuracy concerning the total count of reservoir water levels l_n.

Finally, reservoir water levels gradients gl_n in ascending order of n were incrementally added and *RMSE* were minimized to determine the optimal number of features. In this case, rainfall r_n and reservoir water levels l_n utilized the optimal features obtained from Figures 7 and 8, as shown in Table 4. Figure 9 displays the results of *RMSE* concerning the total count of reservoir water level gradients gl_n. Unlike before the earthquake, a reduction in errors due to reservoir water level gradients was observed after the earthquake, suggesting an increased influence of reservoir water level gradients on seepage following the earthquake. Table 5 presents the optimal number of features for each feature type in each case. It was observed that the optimal number of features for the seepage change was fewer than for the seepage rate. This result suggests that examining changes in seepage reduces the impact on long-term seepage, leading to a decrease in the required number of features.

Table 4. Features used in reservoir water level gradients gl_n evaluation.

	Case X	Case X'	Case Y	Case Y'
r_n	$r_1 \sim r_5$	$r_1 \sim r_2$	$r_1 \sim r_{11}$	$r_1 \sim r_2$
l_n	$l_0 \sim l_{14}$	-	l_0	$l_0 \sim l_1$

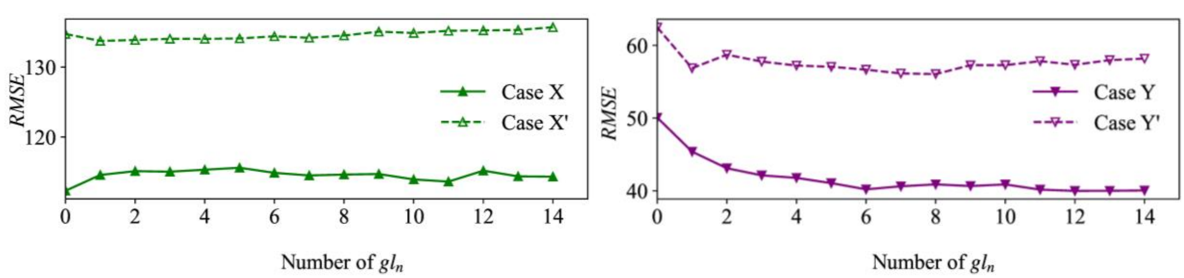

Figure 9. Estimation accuracy concerning the total count of reservoir water level gradients gl_n.

Table 5. The optimal number of features.

	Case X	Case X'	Case Y	Case Y'
r_n	5	2	11	2
l_n	15	0	1	2
gl_n	0	1	12	8

3.4. Estimation Results

Taking practical management into consideration, the seepage rate for the most recent data has been estimated by optimal features obtained via cross-validation. In the estimation procedure for practical management, as depicted in Figure 10, the data was arranged in chronological order, and for each test data, older data than the test data were used as training data. This process was repeated to estimate the most recent 365 days. For the estimation based on the seepage change, the estimated seepage was determined by adding the observed seepage from one day prior to the estimated value. Figure 11 shows the estimated results before the earthquake based on seepage, while Figure 12 shows the estimated results after the earthquake. In the estimation based on the seepage rate, there was a tendency for overestimation and an inability to estimate peaks caused by abrupt seepage changes. On the other hand, in the estimation based on the seepage change, a similar trend to the observed values was observed. Since the estimation based on the seepage change is determined by adding the observed seepage from one day prior to the estimated value, the influence of the seepage from the previous day is significant. Hence, it is considered to be more accurately estimated than the seepage rate, and an improvement in estimation accuracy can be expected by adding the one-day prior seepage as a feature.

By comparing the obtained estimated results with the actually observed seepage data, if differences persist over an extended period, it suggests the possibility of some anomaly. For the targeted Dam A in this study, there are no significant deviations between observed values and estimated values over an extended period. Additionally, there are no unnatural points in the response of seepage during rainfall in the observed data. Therefore, it can be considered as healthy.

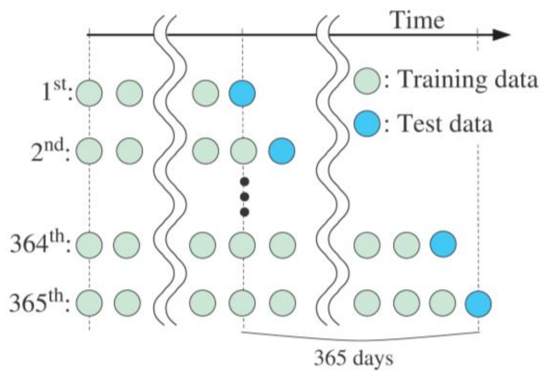

Figure 10. Estimation method in practical management.

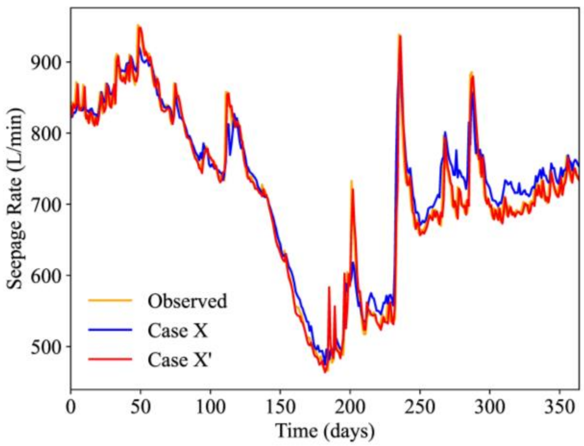

Figure 11. Estimated results before the earthquake.

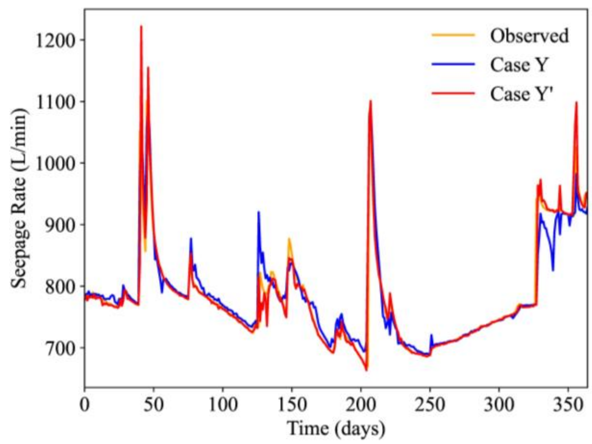

Figure 12. Estimated results after the earthquake.

4. Conclusion

The purpose of this study is to develop a machine learning model for predicting the seepage rate of a dam, leveraging monitoring data to assess the health of the dam. To evaluate the practicality of machine learning models using extremely randomized trees (ERT), estimations of seepage were conducted in Dam A, a rock-fill dam affected by an earthquake, and the optimal number of features was examined. Through the feature selection process, it was revealed that, in Dam A, a 2-day rainfall period had an impact regardless of seismic activity, and the influence of reservoir water level gradients increased due to the earthquake. Additionally, when estimating the seepage change as the target variable, a reduction in the optimal number of features and an improvement in estimation accuracy were observed. Consequently, by examining the disparities between the estimated seepage and observed values, it became possible to detect anomalies in the dam's behavior in real-world scenarios.

5. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grand Number JP23H02327 and JP22K19222. The authors would like to thank the Rural Development Bureau, the Ministry of Agriculture, Forestry and Fisheries for allowing to use the observation data.

6. REFERENCES

Belmokre, A., Mihoubi, M.K., and Santillan, D. (2019). "Seepage and dam deformation analyses with statistical models: support vector regression machine and random forest.'' *Procedia Structural Integrity*, 17, 698-703.

Calamak, M., and Yanmaz A.M. (2014). "Probabilistic assessment of slope stability for earth-fill dams having random soil parameters.'' *Proceedings of the 5th IAHR International Symposium on Hydraulic Structures 2014,* 1-9.

Chen, S., Gu, C., Lin, C., Wang, Y., and Hariri-Ardebili, A. M. (2020). "Prediction, monitoring, and interpretation of dam leakage flow via adaptative kernel extreme learning machine." *Measurement*, 166(15), 1-17.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). "Extremely randomized trees." *Machine Learning*, 63, 3-42.

Hipni, A., El-shafie, A., Najah, A., Abdul Karim, O., Hussain, A., and Mukhlisin, M. (2013). "Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS) ." *Water Resources Management*, 27, 3803-3823.

ICOLD (International Commission on Large Dams) (2017). "Dam safety management: Operational phase of the dam life cycle." ICOLD Bulletin 154, Paris.

ICOLD (International Commission on Large Dams) (2001). “Tailings dams risk of dangerous occurrences: lessons learnt from practical experiences.” ICOLD Bulletin 121, Paris.

Inoue, K., and Suzuki, M. (2020). “Prediction of dam leakage using machine learning.” *Proceedings of the 7th Asian-Pacific Symposium on Structural Reliability and Its Applications*, 341-346.

Li, X., Wen, Z., and Su H. (2021). “An approach using random forest intelligent algorithm to construct a monitoring model for dam safety.” *Engineering with Computers*, 37, 39-56.

Louppe, G. (2014). “Understanding random forests: From theory to practice.” *arXiv:1407.7502*.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., and Kallel, A. (2020). “A machine-learning framework for predicting multiple air pollutants’ concentrations via multi-target regression and feature selection.” *Science of the Total Environment*, 715.

Roushangar, K., and Alipour, S.M. (2018). “Prediction of overland flow resistance and its components based on flow characteristics using support vector machine.” *Water Supply*, 18(4), 1234-1251.

Siacara, A.T., Napa-García, G.F. Beck, A.T., and Futai, M.M. (2020). “Reliability analysis of earth dams using direct coupling.” *Journal of Rock Mechanics and Geotechnical Engineering*, 12(2), 366-380.

Sorkhabi, O.M., Shadmanfar, B., and Kiani, E. (2022). “Monitoring of dam reservoir storage with multiple satellite sensors and artificial intelligence.” *Results in Engineering*, 16, 100542.

UNESCO World Water Assessment Programme (2012). “Facts and figures; from the United Nations world water development report 4: managing water under uncertainty and risk.” The United Nations World Water Assessment Programme.

Wang, W-C., Chau, K.-W., Cheng, C.-T., and Qin, L., (2009). “A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series.” *Journal of Hydrology*, 374, 294-306.