The Seventh Asian-Pacific Symposium on Structural Reliability and Its Applications (APSSRA2020)
October 4–7 2020, Tokyo, Japan
T. Takada, I. Yoshida & T. Itoi (editors)

Prediction of Dam Leakage Using Machine Learning

K. Inoue[1] and M. Suzuki[2]

[1]*Graduate School of Agricultural Science, Kobe University. Email: mornel@kobe-u.ac.jp*
[2] *Graduate School of Agricultural Science, Kobe University. Email: msuzuki@peridot.kobe-u.ac.jp*

Abstract: This paper describes the prediction methodology of dam leakage using machine learning algorithms including random forest (RF), extremely randomized trees (ERT) and support vector regression (SVR). In this study, observation data monitored at an actual rock-fill dam located at Kyushu province in Japan were used in order to build the training data and testing data, which include data set related to the water level in the dam reservoir and rainfall depth. The predicted and observed leakages not only for the filling duration in nine months but also for the all observation data in four years were compared, demonstrating that they were in good agreement with each other. Comparative study was extended to exhibit the superior performance of ERT to other algorithms, which comprises an ensemble of unpruned decision trees according to an information criterion which is maximized by one of the splitters. The adaptive boosting, which is one of the boosting algorithms and aims to convert a set of weak classifiers into a strong classifier, was used to assess the feature importance of input variables to the leakage prediction. As a result, it was revealed that the water level in the reservoir, the temporal gradient of water level and the rainfall a day are of importance. Also, the rainfall two days before was classified as a relatively importance variance among rainfall data observed one to seven days before.

Keywords: machine learning, extremely randomized trees, dam leakage, prediction.

1. Introduction

Although the amounts of surface and subsurface water resources are enough for the entire world, spatial and temporal distribution of these resources exhibits uneven pattern. Water for irrigation and food production constitutes one of the greatest pressures on freshwater resources. Agricultural activity accounts for approximately 70% of global freshwater withdrawals (UNESCO World Water Programme, 2012). Water need is steadily increasing around the world, even though the freshwater resources are limited and unevenly allocated. Also, to provide more clean energy, hydroelectric developments of small or large size, whether run of the river or of accumulated storage, fit the concept of renewable energy.

Dams which provide regular water from dam reservoirs based on a demand pattern are a vital part of the civilization. Dams have played a significant role for a long time in human life and natural systems, since the ancient innovation of gravity dam and earth dam in ancient Egypt around 2950-2750 B.C. and in Mesopotamia around 2000 B.C., respectively. In addition to water supply, flood control and irrigation, a variety of functional priorities such as power generation and industrial water supply have been incorporated into the original design and construction of a dam. After the built of a dam, its structure and component parts begin to age to a greater or lesser extent. The unique nature of each dam means that every part comprising the dam gradually deteriorates at a different rate in a different way. Some of the dams may remain safe for a thousand years, others may have cracks and leakage after less than a decade. Whereas approximately 2,700 dams exist in Japan, the Ministry of Agriculture, Forestry and Fisheries in Japan has direct control over the dams of 189 in which approximately half of the dams are now more than 30 years old. The International Committee on Large Dams pointed out that in the future attention and activity will be more and more shifted from the design and construction of new dams to the restoration of the structural and operational safety of existing dams (ICOLD, 2017). ICOLD (2017) also advocated a few key concepts, for example, ``an effective dam safety management program must address interrelationships amongst technical and management aspects of program activities in all life cycle stages''.

As well as other civil infrastructures, monitoring of dams focuses on a different area of the dam body, on the relevant components and on physical processes involving structural deformations, water infiltration and corrosion. Basic monitoring data are the water level of a reservoir and a phreatic surface and surface deformation of a dam body to prevent from flooding and to contribute the likelihood of the slope stability (Calamak and Yanmaz, 2014, Siacara et al., 2020). Additionally, the seepage discharge flow from the filter drains (ICOLD, 2001) is also vital information measured at a site. In dam management, especially in earth-fill or rock-fill dams, seepage rate through, below, or around dams is an essential indicator of the health and condition of the dam. The transitional amount of seepage, or leakage, is generally associated with the water level in the reservoir. However, any abrupt change in the amount of leakage may be a serious sign of deterioration of dam body and indicate serious problems that threaten the immediate safely of the dam. Therefore, the ability to adequately predict an anomalous behavior of dam seepage rate can be crucial to reduce the risk and the likelihood of dam failure. Probable emergencies may include clogging of the spillway, development of internal erosion, resulting in repairs to spillways, embankments, or other dam appurtenant structure.

In recent years, soft computational approaches are being applied to solve the practical problems and to identify the optimal values in different scientific disciplines (Masmoudi et al., 2020). Approaches using artificial intelligence (AI) or machine learning (ML) are remarkable forecasting tools that have been successfully applied to several problems including dam engineering (e.g. Wang et al., 2009, Hipni et al., 2013, Roushangar and Alipour, 2018). Several methodologies are applied to

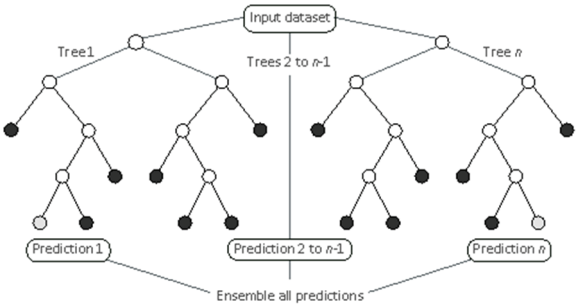

Figure 1. Structure of random forest.

problems including support vector machine (SVM) and random forest (RF), which are representatives of the widely used branches of soft computing approaches. Surprisingly, there was limited information on the prediction of seepage discharge rate using these methodologies, indicating the possibility of the application of other machine learning approaches to dam management problems. The aims of this study are to develop the prediction means associated with dam seepage rate using machine learning and to compare a few machine learning methods based on the transitional data measured at the actual rock-fill dam.

2. Materials and Methods

2.1 Study area and observation data

In this study, the raw data of 32,485 observations of seepage rate were collected from the filter drains at the interval of one hour at X-dam having over 50 m high from 2014 to 2017. The X-dam is a rock-fill embankment dam, capable of storing more than eight million cubic meters, serves mainly for irrigation and is located at the southern part of Kyushu province in Japan. The dam was completed in 2015 after filling the reservoir in 2014.

A measurement facility for seepage rate is located at the downstream side apart from the toe of dam and measures the discharge one per hour automatically. Here 32,485 seepage rate data sets were preprocessed after data clean such as the remove of data sets with missing data. In addition to the dataset of seepage rate, water level and rainfall data sets were also obtained and mainly utilized as training and test data set in the prediction of seepage rate.

2.2 Random forest

In machine learning approaches, an ensemble method is a technique that combines the decisions from multiple machine learning models together to make more accurate predictions than any individual model. There are two types of the ensemble learning such as boosting and bagging. Boosting refers to a family of algorithms which converts a weak learner to a strong learner, while bagging is the application to reduce the variance for ensemble algorithms having high variance.

This study applied random forest regression (RF) to model the seepage rate measured at X-dam site. Random forest regressor is a non-parametric machine learning technique that has been developed for multi-class classification and regression problems and is a type of ensemble machine learning algorithm classified as bagging (Breiman, 2001). The regression problem is solved by combining predictions from a large number of individual trees and using bootstrapped samples drawn from the original learning sample. Random forest structure is shown in Figure 1. Random forest has some advantages such as handling thousands of input variables without variable deletion and maintaining its accuracy under the missing of a large proportion of the data.

Based on the training data set, random forest learns linkages between the covariance and the target variable at sampled locations (Koch et al., 2019). Two important parameters of the random forest algorithm are the number of maximum depth of the trees to grow, and the number of trees in the forest to build before taking the maximum voting or averages of predictions. Also, complexity parameter used for minimal cost-complexity pruning may contribute the accuracy of machine learning. Here, we estimate the seepage rate through, below, or around dams with the random forest regression where the water level in the reservoir and the rainfall depth are utilized as input data. Parameters related to the number of trees and the maximum depth of the tree are set to 50 and 43, respectively, based on the grid search described later.

2.3 Extremely randomized trees

Extremely randomized trees (ERT) is a tree-based ensemble learning algorithm for supervised classification and regression problems and was proposed by Geurts et al. (2006) and is also called as extra trees. Extremely randomized trees aggregates the predictions of the decision trees to significantly reduce the computational complexities, splits nodes by randomly selecting cut-points and utilizes all training samples in order to grow to trees instead of the bootstrap approach. The splitting procedure for regression problems in extremely randomized trees has two main parameters including the number of random splits at each node and the minimum sample size for splitting a node (Seyyerdattar et al., 2020).

As well as ensemble averaging, the randomization of the input features and cut points lead to the reduction of the variance. This model efficiently solves variance problems and mines more valuable information compared to other widely used tree-based approaches such as the decision tree and random forest aforementioned before. Also, remarkable features associated with the scalability and consistency property of tree-based supervised learning allow to apply to complex and large scale problems with high dimensionality or non-linearity (Wehenkel et al., 2006). Like the random forest, parameters related to the number of trees and the maximum depth of the tree are set to 50 and 43, respectively, based on the grid search described later.

2.4 Support vector regression

Support vector regression (SVR) is a type of support vector machine, which aims to achieve a linear and non-linear regression (Vapnik, 1995). The ultimate goal of the support vector regression is to find a regression function based on the training data.

$$f(x) = wX(x) + B \tag{1}$$

where $f(x)$ is the regression function, or the objective variable, $X(x)$ is a function based on the explanatory variables x, w is the regression coefficients, and B is the residues. The problem to be solved is translated into the minimization (or maximization) problem of the objective function with the use of relaxation factors.

$$minimize \ \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

$$f(x) \leq wX(x) + \varepsilon + \xi_i \tag{3}$$

$$f(x) \geq wX(x) - \varepsilon - \xi_i^* \tag{4}$$

where ε is the error, C is the error penalty parameter, and ξ_i and ξ_i^* are the relaxation factors and are greater than 0. The error penalty parameter means the extent of the penalty for a sample out of the error ε and affects the tradeoff between the computational complexity and the degree of mistakenly classified samples (Solgi et al., 2017). The parameters ε and C were set to 0.1 and 1.0, respectively.

In the support vector regression, some kernel functions, which are linear, polynomial, sigmoid and Gauss kernel function, have an effect on the computational accuracy. Gauss kernel function, which is called as the radial basis function representing two samples as feature vectors in some input space, was adopted in this study.

3. Results and Discussion

3.1 Training and testing data used in this study

As for the seepage rate prediction of the X-dam, this study utilized 32,485 observation data measured at one-hour interval associated with the water level and the rainfall from February 1st, 2014 to October 31st, 2017. In addition to these two measured data, 8 types of the rainfall data such as the rainfall half a day before and one day through 7 days before were employed since the seepage rate may be influenced not only by the rainfall event of the day but also by the rainfall event before. Moreover, the gradient of the water levels between the day and a day before was used as another variable to achieve the prediction of the seepage rate. Thus, 11 variables, each of which has 32,485 observation data, were used in this study.

Each variable was standardized using their respective Z-score prior to inputting training models so as to keep all the 11 variables possessing the same degree of influence on ultimate seepage rate prediction. After reviewing the all clean data measured 32,485 observations, 6,071 data observed during the first filling to investigate the condition of the dam conducted from February 1st, 2014 to October 14th, 2014 were extracted. The data were separated into a training set of 75% and a testing data set of 25%.

3.2 Seepage rate prediction using filling data

Seepage rate prediction results by random forest (RF) and extra trees (ERT) are shown in Figures 2 and 3, respectively. The upper graph depicts the observed transitional data of rainfall and water level in the X-dam reservoir, while the middle shows the observed seepage rate. The lower graph exhibits the results in the testing phase. Seemingly, in both figures, testing phases are good agreement with the observation data whereas the differences between the observed and predicted results appear at some points expressing the abrupt increase of seepage rate, especially for the case predicted using random forest.

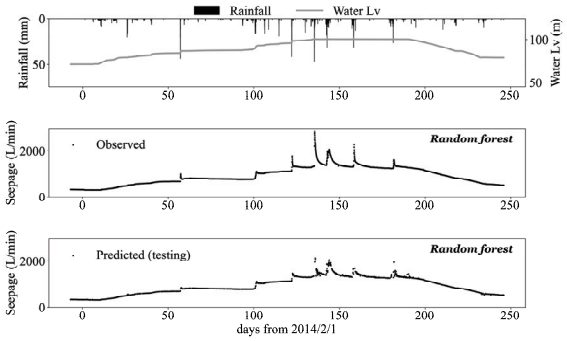

Figure 2. Prediction results of the seepage rate by random forest: (upper) the transitional data of rainfall and water level in the dam reservoir, (middle) observation data of seepage rate and (lower) results in the testing phase.

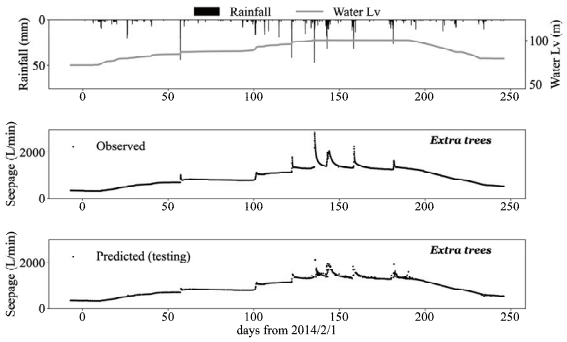

Figure 3. Prediction results of the seepage rate by extra trees: (upper) the transitional data of rainfall and water level in the dam reservoir, (middle) observation data of seepage rate and (lower) results in the testing phase.

Like Figures 2 and 3, Figure 4 shows the results of the seepage rate predicted by support vector regression (SVR). As it can be seen, SVR is not able to capture the transitional variation of the seepage rate. SVR provides the average values of the seepage rate as a whole, which may the nature of SVR separating the characteristics of data variation of concern. The performance of the predictive three algorithms employed here was evaluated in terms of the mean squared error (*MSE*) and mean absolute error (*MAE*) defined by the following equations:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Obs_i - Pred_i)^2 \tag{5}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\|Obs_i - Pred_i\| \tag{6}$$

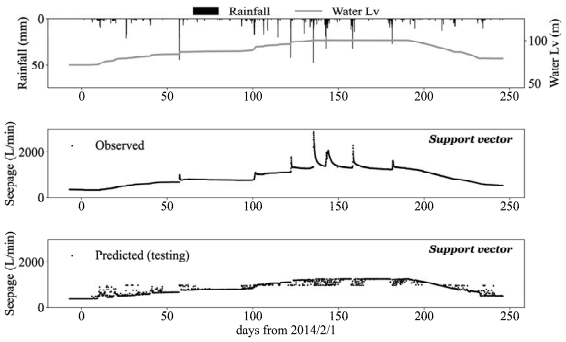

Figure 4. Prediction results of the seepage rate by support vector regression: (upper) the transitional data of rainfall and water level in the dam reservoir, (middle) observation data of seepage rate and (lower) results in the testing phase.

In equations (5) and (6) to assess the prediction accuracy of the algorithms, N is the number of data, Obs_i and $Pred_i$ are the ith measured and predicted values, respectively. Table 1 shows the statistical metrics for algorithms' performance evaluation in the training and testing phases. In comparison to the RF and SVR, the performance of ERT in both phases was far better than those of other algorithms. The primary advantage and the verified accuracy of ERT over the other algorithms for regression purposes are found in the report of Seyyedattar et al. (2020). Therefore, ERT was adopted as the computational algorithm for predicting seepage rate in this study.

3.3 Feature importance

Before investigating the importance of features of concern, the effect of two parameters in ERT on the accuracy were presented using the grid search, which is the process of scanning the data to configure optimal parameters for a certain algorithm. In order to find best combination, the range of the number of trees, which is expressed as the number of estimators, and the maximum depth of trees were set to 50 to 350 and 10 to 70, respectively, at the unit interval. Base on the all combinations of the number of estimators and the maximum depth of trees, values of *MAE* were computed using ERT. Figure 5 shows the results of grid search with the distribution map of *MAE*, demonstrating that the best combination of the number of estimators and the maximum depth of trees are 50 and 43 in ERT.

As aforementioned above, eleven variables which are mainly derived from the rainfall data were initially prepared and considered as factors relevant to seepage rate. Despite of the regression problems, in many cases it is quite important to not only have an accurate but also a contribution of variables. Knowing feature importance indicated by machine learning is of significant concern to acquire a better understanding of the model's logic and a step stone for improving the model by focusing on some important variables.

Table 1. Performance evaluation of three regressors.

Algorithm	Training data		Testing data	
	MAE	*MSE*	*MAE*	*MSE*
RF	5.25	472.53	10.54	1962.78
ERT	1.39	43.51	11.0	2041.07
SVR	95.87	3386.165	97.24	32780.65

* *MAE*: mean absolute error, *MSE*: mean squared error

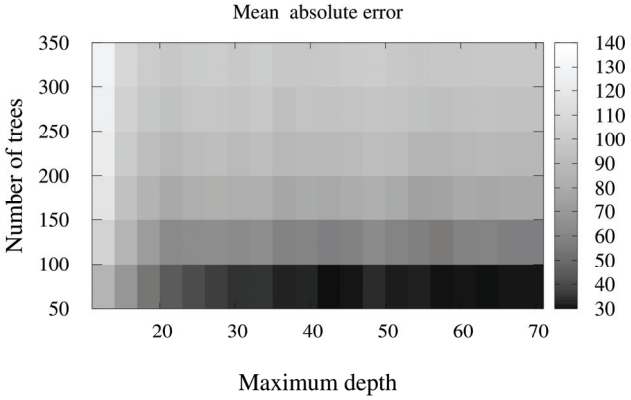

Figure 5. Results of grid search in ERT.

In this study, the adaptive boosting, which was proposed by Freund and Shapire (1997) and is called as AdaBoost algorithm, was employed for generating a strong classifier from a set of weak classifiers and for assessing the feature importance of eleven variables. AdaBoost algorithm creates a set of poor learners by maintaining a collection of weights over training data and adjusts them after each weak learning cycle adaptively. The weights of the training samples which are misclassified by current weak learner will be increased while the weights of the samples which are correctly classified will be decreased (Li et al., 2005). The results of feature importance by AdaBoost are shown in Figure 6. In this figure, WL means the water level, Gw stands for the gradient of water level in a day, r05 is the rainfall half a day before and ``rx'' means the rainfall x days before. The water level in the reservoir is the most importance factor for seepage rate prediction followed by the gradient of water level and the rainfall for a day. Surprising though it may seem is that the rainfall 2 days before turned out to be more important than other past rainfall data. This outcome may indicate the outflow characteristics of the surrounding mountains of the X-dam.

Although all the selected variables are related to seepage rate, the less importance variables my cause noise and reduce the prediction accuracy. To examine the most effective combination of variables, the features were eliminated two by two starting from the least important factors shown in Figure 6. Through the computation using ERT under the reduction of the number of features, the relation between the number of features and the results of statistical metrics MAE and MSE is exhibited in Figure 7. The highest performance was achieved when all of the eleven variables were employed. This is attributed to the

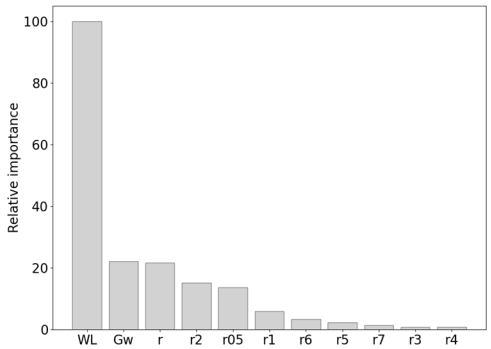

Figure 6. Results of feature importance using AdaBoost.

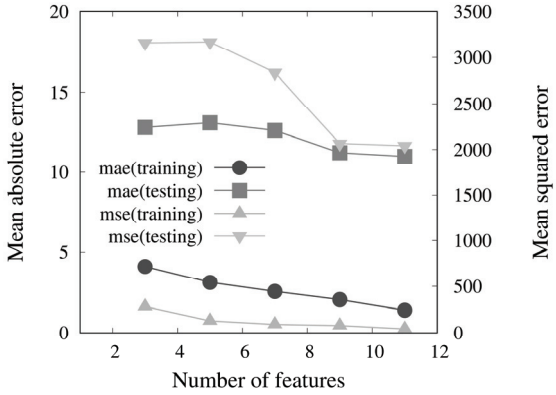

Figure 7. Relation between the number of features and statistical metrics.

fact that the total number of observation data, 6,071 data, may reflect the low number of appropriate data for predicting the seepage rate in machine learning.

3.4 Seepage rate prediction using full data

The machine learning algorithm of ERT using all observation data of 32,485 was conducted to predict the seepage rate in about four years. Figure 8 shows the prediction results with the observed data of rainfall and water level in the X-dam reservoir in testing phase. In the middle and lower graphs, the plots of observed and predicted seepage rate variations were shown, respectively, demonstrating the good agreement with the observation data with the values of 3349.9 and 21.2 of *MSE* and *MAE*, respectively. The value of squared *MSE* divided by *MAE* becomes 2.7 and expresses the improvement of accuracy from 4.1 computed from the data in Table 1.

The scatter plot of the predicted versus observed seepage rate are depicted in Figure 9 where the results of the training and testing phases are plotted. It is observed that the scatter points of the testing results are relatively good distribution with less variance around the regression line. Dams are subjected to the daily water level fluctuation such as rapid drawdown and refill, resulting in induction of a structural impact on the behavior of dam body. This methodology yields meaningful information for distinguishing the stable behavior and unstable behavior of dam leakage and for the reduction of disaster risk.

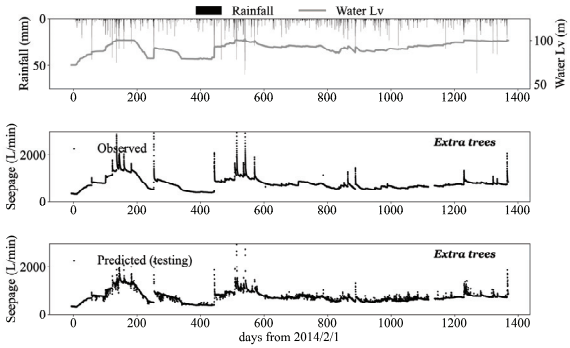

Figure 8. Prediction results of the seepage rate by extra trees: (upper) the transitional data of rainfall and water level in the dam reservoir, (middle) observation data of seepage rate and (lower) results in the testing phase.

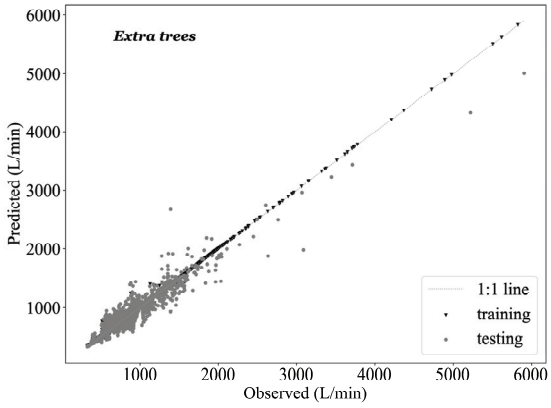

Figure 9. Plot of observed data versus predicted results using extra trees in training and testing phases.

3.5 Comparison with other regressors

In machine learning, a wide variety of algorithms exist, indicating that selection of an algorithm may be required according to the problem of interest. For the comprehensive comparing the seepage rate prediction performance among a few algorithms, AdaBoost, Bagging and Gradient boosting as well as Random forest and Extremely randomized trees, were selected and conducted the prediction using full observation data. The performance of each algorithm was assessed using the correlation coefficient *R*:

$$R = \frac{\sum_{i=1}^{N}(Obs_i - Pred_a)(Pred_i - Pred_a)}{\sqrt{\sum_{i=1}^{N}(Obs_i - Obs_a)^2 (Pred_i - Pred_a)^2}} \quad (7)$$

where Obs_a and $Pred_a$ are the mean measured and predicted values, respectively. Table 2 lists the results of scores for each regressor in training and testing phases. It

Table 2. Comparison of the results of regressors.

Algorithm	Training score	Testing score
AdaBoost	0.9775	0.8889
Bagging	0.9772	0.9027
Bagging & AdaBoost	0.9360	0.8902
Gradient boosting	0.8798	0.8479
Random forest	0.9774	0.9034
Extremely randomized trees	0.9850	0.9192

is revealed that ERT outperforms other algorithms. This may be attributed that this tree-based method splitting nodes and using whole data has the additional feature of bounded input-output approximation in the iterative fitting procedure (Wehenkel et al., 2006).

4. Conclusions

Dam leakage prediction is crucial for strict dam-use planning and disaster risk reduction in a dam site. In this study, using the observation data of the water level in the X-dam's reservoir located in Kyushu province in Japan and the rainfall depth, seepage rate prediction was conducted. Three machine learning algorithms such as random forest (RF), extremely randomized trees (ERT) and support vector regression (SVR) were compared. ERT which splits nodes by choosing cut-points fully at random and utilizes the whole learning sample to grow the trees showed a high performance. The predicted and observed leakages not only for the filling duration in nine months but also for the all observation data in five years were compared, demonstrating that they were in good agreement with each other. Moreover, it was revealed that the water level in the reservoir, the temporal gradient of water level and the rainfall a day are of importance. ERT algorithm is highly recommended to conduct the prediction of dam leakage, whereas a more accurate prediction should be achieved through the increase of observation data or a minor change of the algorithm.

Acknowledgments

This work was supported by JSPS Grant-in-Aid for Scientific Research (B) Grand Number JP19H03074. The authors would like to thank the Rural Development Bureau, the Ministry of Agriculture, Forestry and Fisheries for allowing to use the observation data.

References

UNESCO World Water Assessment Programme 2012, Facts and figures; from the United Nations world water development report 4: managing water under uncertainty and risk, 16p.

ICOLD (International Commission on Large Dams) 2017. Dam safety management: Operational phase of the dam life cycle. In *ICOLD Bulletin 154*, 240p.

Calamak, M. and Yanmaz A.M. 2014. Probabilistic assessment of slope stability for earth-fill dams having random soil parameters. In *Proceedings of the 5th international symposium on hydraulic structures 2*014. https://doi.org/10.14264/uql.2014.16.

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), pp.366-380.

ICOLD (International Commission on Large Dams) 2001. Tailings dams risk of dangerous occurrences: lessons learnt from practical experiences. In *ICOLD Bulletin 121*, 144p.

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.

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, J*ournal of Hydrology*, 374, pp.294-306.

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, pp.3803-3823.

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), pp.1234-1251.

Breiman, L. 2001. Random forests, *Machine Learning*, 45, pp.5-32.

Koch, J., Berger, H., Henriksen, H.J. and Sonnenborg, T.O. 2019. Modelling of the shallow water table at high spatial resolution using random forest, *Hydrology and Earth System Sciences*, 23, pp.4603-4619.

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

Seyyedattar, M., Ghiasi, M.M., Zendehboudi, S. and Butt, S. 2020. Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models, *Fuel*, 269, 116834. https://doi.org/10.1016/j.fuel.2019116834

Wehenkel, L., Ernst, D. and Geurts, P. 2006. Ensembles of extremely randomized trees and some generic application, In *Robust methods for power system state estimation and load forecasting*, *Versailles, France*.

Vapnik, V. 1995. *The nature of statistical learning theory*, Splinger, New York.

Solgi, M., Najib, T., Ahmadnejad, S. and Nasernejad, B. 2017. Synthesis and characterization of novel activated carbon from Medlar seed for chromium removal: Experimental analysis and modeling with artificial neural network and support vector regression, *Resource-Efficient Technologies*, 3, pp.236-248.

Freund, Y. and Shapire, R. 1997. A decision-theoretic generalization of on-line learning and application to boosting, *Journal of Computer and System Sciences*, 55, pp.119-139.

Li, X., Wang, L. and Sung, E. 2005. A study of AdaBoost with SVM based weak learners, In *Proceedings of International Joint Conference on Neural Network*, pp.196-201.