Hindawi
Advances in Meteorology
Volume 2018, Article ID 5024930, 11 pages
https://doi.org/10.1155/2018/5024930

Research Article

Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data

Changhyun Choi,[1] Jeonghwan Kim,[2] Jongsung Kim,[1] Donghyun Kim,[1] Younghye Bae,[2] and Hung Soo Kim[1]

[1]*Department of Civil Engineering, Inha University, Incheon 22212, Republic of Korea*
[2]*Institute of Water Resources System, Inha University, Incheon 22212, Republic of Korea*

Correspondence should be addressed to Hung Soo Kim; sookim@inha.ac.kr

Received 13 February 2018; Accepted 15 May 2018; Published 13 June 2018

Academic Editor: Alastair Williams

Prediction models of heavy rain damage using machine learning based on big data were developed for the Seoul Capital Area in the Republic of Korea. We used data on the occurrence of heavy rain damage from 1994 to 2015 as dependent variables and weather big data as explanatory variables. The model was developed by applying machine learning techniques such as decision trees, bagging, random forests, and boosting. As a result of evaluating the prediction performance of each model, the AUC value of the boosting model using meteorological data from the past 1 to 4 days was the highest at 95.87% and was selected as the final model. By using the prediction model developed in this study to predict the occurrence of heavy rain damage for each administrative region, we can greatly reduce the damage through proactive disaster management.

1. Introduction

The occurrence of natural disasters such as floods, tsunamis, and earthquakes is increasing due to the climate change. Also, the damage is becoming larger and larger due to the rapid urbanization over the world. In South Korea, about 65% of all damage is due to heavy rain, and thus there is a pressing need for countermeasures [1]. If the scale and impact of such damage is estimated quickly in advance, this makes disaster management more possible at the preventive and preparatory stages, and this would help to avoid large-scale damage due to heavy rain like that which occurred in Hongcheon and Cheongju in the summer of 2017. In particular, if there is rapid predisaster forecasting of expected damage by the administrative division for the regions that will be affected, this can be of great help to policymakers in setting up and implementing disaster prevention measures. Moreover, it will be possible to establish a voluntary disaster management system in which citizens themselves can prepare for disasters and expected damage by receiving forecasts about them.

Previous studies that were used in predicting and preparing for natural disaster damage in advance mostly performed linear regression analysis using weather factors such as precipitation, rainfall intensity, maximum wind speed, and hurricane central pressure that cause natural disasters and damage through floods, rainstorms, and hurricanes [2–11]. These studies analyzed the relationship between weather factors and damage extent through regression analysis, and they used the constructed regression models to attempt to predict the extent of damage through weather factors alone. However, it proved difficult for most of these models to predict the actual extent of damage adequately. In order to overcome the shortcomings of such studies, others have taken into account socioeconomic factors such as per capita income, population density, and imperviousness of an area in addition to weather factors that directly give rise to natural disasters [12–16]. Although the inclusion of socioeconomic factors besides weather factors led to some improvement in the prediction performance of these linear regression models, the nonlinear character of disasters and their damage scale present problems that cannot be solved by them. More recently, rapid advances in computing technology and data processing speed have led to the emergence of studies that apply big data and machine learning to disaster management [17, 18]. The predominant approach in all these studies is to use just a handful of explanatory variables in a regression model to estimate the damage scale of disasters. In regard to disaster management research in Korea, there is in particular a dearth of studies that use machine learning, which is known to be able to maximize the prediction performance of models, and big data, which produce valuable information through various data that could not previously be taken into account.

Accordingly, the present study relies on the meteorological big data provided by the Korea Meteorological Administration to arrive at a list of various explanatory variables that account for the occurrence of heavy rain damage and uses machine learning—known to have higher prediction performance than regression models—to develop functions that can predict heavy rain damage in advance. For this purpose, we constructed a response variable and explanatory variables for the study area of our study and used various machine learning models such as decision trees, bagging, random forests, and boosting to develop prediction models for heavy rain damage based on big data. We used two algorithms in developing the prediction functions, namely, Algorithm 1 that uses same-day weather observation data to make predictions and Algorithm 2 that uses past weather observation to do so. Models were constructed on this basis, and we thereby developed a prediction model for heavy rain damage that can be used immediately in actual practice.

2. Theoretical Background

2.1. Machine Learning. Machine learning is a field concerned with deriving new knowledge by feeding the requisite data to a computer and making it learn from them like a human being studying a new subject area. For example, suppose that there is a set of pairs (x, y) with the data (1, 7), (2, 14), (3, 21), and (5, 35) already given as members of the set. Even if a computer does not know the function for y, machine learning can be used to make it provide, say, the y values for (7, ?) or (10, ?) after the data are entered and the computer learns from them. That is, the computer will give the answers even without directly programing it with the function $y = 7x$. In machine learning, there are two main types of learning method. One method is supervised learning that is used to infer the function for y, and the other is unsupervised learning that is used to determine how the data for x values are distributed. The present study uses decision tree learning, which is a representative technique in machine learning, along with ensemble methods based on decision tree models such as bagging, random forests, and boosting, in order to develop a prediction model for heavy rain damage. All the methods used here are supervised learning techniques, which use their own algorithms to generate rules that best explain the response variables.

2.2. Decision Tree Models. Decision tree models can be used in both classification and regression, and they express results in the form of tree-shaped graphs. A decision tree finds rules that best explain values of a response variable by recursively partitioning the space of each explanatory variable. If the entire domain of explanatory variables is partitioned into M number of domains $R_1, \ldots, R_M$ on the basis of the criterion minimizing the classification error rate, then the Gini index G and cross-entropy E are mainly used as related criteria to determine this, as shown in (1):

$$\begin{aligned} G &= \sum_{m=1}^{M} p_{mk}\left(1 - p_{mk}\right), \\ E &= -\sum_{m=1}^{M} p_{mk} \cdot \log\left(p_{mk}\right). \end{aligned} \tag{1}$$

In (1), p_{mk} indicates the proportion of the data in the mth partition that belongs to class k of the response variable. The response variable in this study has two classes, 1 and 0, and thus k has the values 1 and 0. A decision tree grows through top-down partitioning. After the first split of the domain of explanatory variables into partitions that minimize the indices given in (1), the resulting partitions are again split into further partitions that minimize the same indices. This goes on until the degree of minimization becomes very minute, or when a prespecified stopping condition is met. For a decision tree that has stopped growing, pruning is automatically performed to prevent overfitting.

In general, a decision tree can have lower prediction performance than other prediction models, but it has the advantage of being relatively easy to interpret. However, if decision trees are actively used in the ensemble techniques described below, this will not only compensate for the weaker prediction performance of a single decision tree, but it can even exhibit equal or greater prediction performance than other complex models. Figure 1 shows a schematic of the decision tree concept.

2.3. Ensemble Methods. An ensemble method constructs multiple prediction models for a given dataset and then combines these models into a final prediction model. The first ensemble algorithm to be proposed was Breiman's [19] bagging, based on the bootstrap method, followed later by boosting based on Freund and Schapire's [20] AdaBoost algorithm. There is also the random forest algorithm proposed by Breiman [21]. Various other ensemble methods have been developed since then, but bagging, boosting, and random forests are the most popular ensemble methods that remain widely in use, and many studies have established that these methods can maximize the prediction performance of models.

Bagging, boosting, and random forests mainly use a single model repeatedly to aggregate the results. The model used is usually the decision tree model explained above. Since decision trees can be applied to classification as well as regression problems, ensemble methods that use decision trees can also be applied to both kinds of problems. While a single decision tree divides the space of explanatory variables into discrete partitions, ensembles using decision trees as base models average or vote over several differently partitioned

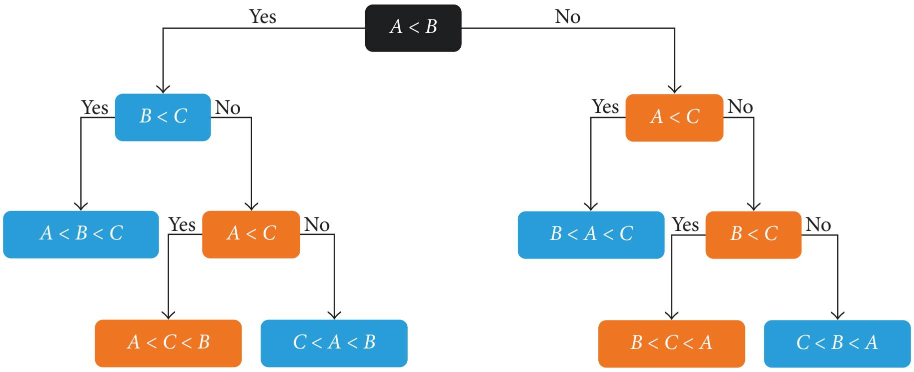


Figure 1: Decision tree concept.

decision trees. Thus, ensembles have the advantage of naturally learning nonlinear effects in addition to linear effects.

2.3.1. Bagging. Bagging generates multiple bootstrap data from the original dataset, constructs a prediction model in a uniform way for each bootstrap data, and combines the models to arrive at the final model. Here, the term "bootstrap data" refers to a dataset obtained by random sampling with a replacement that has the same size as the original dataset. More formally, let us refer to the original dataset as $D = (X, Y)$, and B number of bootstrap datasets as $D^{(b)} = (X^{(b)}, Y^{(b)})$, $b = 1, \ldots, B$. Then, for each bootstrap dataset $D^{(b)} = (X^{(b)}, Y^{(b)})$, we construct model $f^{(b)}(x)$. This yields B number of prediction models whose results in regard to a classification problem can be combined, as shown in the following equation:

$$f(x) = \arg\max_k \left[\sum_{b=1}^{B} I\left(f^{(b)}(x) = k\right) \right]. \qquad (2)$$

Equation (2) involves voting on the results of the B number of prediction models. This indicates that if the majority of these models predict class k for a response variable, then class k will be decided as the final prediction result. Figure 2 shows a schematic of the bagging concept.

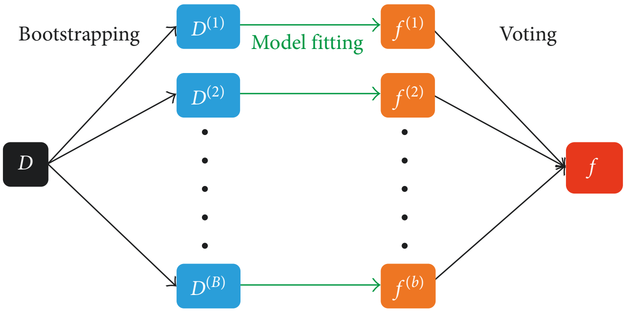


Figure 2: Bagging concept.

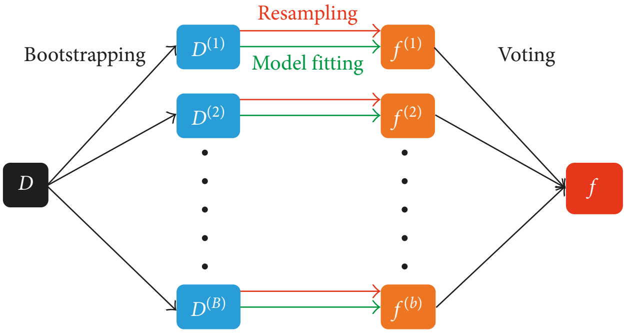


Figure 3: Random forest concept.

2.3.2. Random Forests. Random forests use almost the same algorithm as bagging, but they differ from the latter in adding random sampling of explanatory variables in the process of generating bootstrap data. In the case of bagging, the models it generates could depend on just a few explanatory variables that are strong predictors, and consequently the predicted values of models in bagging can become highly correlated with one another, thus posing the risk of leading to higher prediction variance than a single model. However, a random forest has the effect of reducing such prediction variance in bagging by solving the problem through the random sampling of explanatory variables. Just like bagging, a random forest also constructs the model $f^{(b)}(x)$ for each of the B number of bootstrap datasets and then generates the final model in the same way as (2) given in Section 2.3.1. Figure 3 shows a schematic of the random forest concept.

2.3.3. Boosting. Boosting is similar to bagging and random forest in generating multiple single models and aggregating their results but differs from them in updating the weights of the observations at each iteration while continuously using the same original dataset. More formally, at the first stage, model $f^{(1)}(x)$ is fitted using the dataset $D = (X, Y)$ and weight assignment $W^{(1)} = (w_1^{(1)}, w_2^{(1)}, \ldots, w_n^{(1)})$, where the weights are adjusted to sum up to 1. Then, the predicted results of model $f^{(1)}(x)$ are compared to the actual values of y, and the weights of well-classified observations are reduced while those of misclassified observations are increased to obtain the updated weight assignment $W^{(2)} = (w_1^{(1)}, w_2^{(2)}, \ldots, w_n^{(2)})$. At the second stage, model $f^{(2)}(x)$ is fitted using the dataset $D = (X, Y)$ and weight assignment $W^{(2)} = (w_1^{(1)}, w_2^{(2)}, \ldots, w_n^{(2)})$. In this manner, B number of prediction models are constructed,

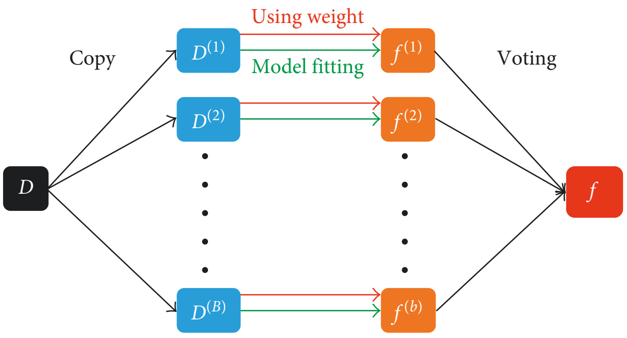


Figure 4: Boosting concept.

and the final model is generated in the same way as (2) given in Section 2.3.1. Figure 4 shows a schematic of the boosting concept.

2.4. Undersampling. The response variable Y used in this study is a binary categorical variable that has two possible values, 1 and 0. A value of 1 means the occurrence of heavy rain damage, and a value of 0 means no damage. However, there is a problem in that the ratio between the two classes is highly asymmetrical. That is, the data are unbalanced, class 1 being sparse and class 0 being major. There are research results indicating that an unbalanced class ratio involving a response variable is detrimental to the performance of classification models [22]. For instance, if the proportion of 0 is 90%, then a model that predicts 0 in all cases might look like a good prediction model because it has an accuracy of 90%, but actually this is a model that has no prediction performance regarding 1. Various sampling techniques have been proposed to improve the performance of binary classification models in regard to such unbalanced data. Among these, the present study uses the undersampling method to develop its models. Undersampling is a method of removing imbalance in the original data by adjusting the size of the major class sample through random sampling to match the size of the minor class. In other words, it involves reducing the number of cases in the 0 class to that of the 1 class so as to convert the unbalanced data to a balanced one. We judged this to be an appropriate strategy to use, given the considerably large size of the data used in this study and the appreciable amount of time required for trial and error in the process of analysis. Figure 5 shows a schematic of the undersampling concept.

2.5. Model Development Process. When developing the prediction model, we constructed it by first distinguishing between training data and test data. After that, a model was constructed using only the training data and then applied to the test data that were not used in the model's training to evaluate its prediction performance objectively. If this evaluation determined that the model can be used in prediction, then both the training data and the test data could be used to update the prediction model. This updated model can yield predictions for the response variable in relation to new sets of explanatory variables in the future.

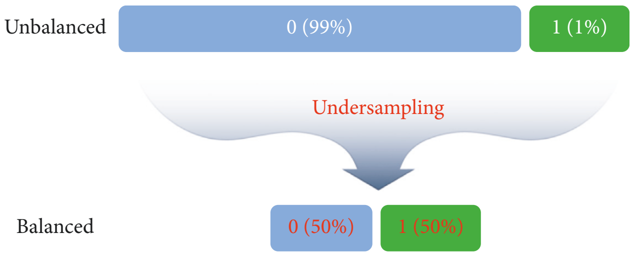


Figure 5: Undersampling concept.

In this study, the model was developed by reducing the size of the data through undersampling in its training stage and then applied to the test data to evaluate its prediction performance. Thus, it was necessary to follow this process of training and evaluation in training a model to find the optimum value for some specific tuning parameter of the model. Therefore, in this study, we implemented the 10-fold cross-validation method, wherein a model was fitted to the data refined through undersampling and then applied to the validation data to test its prediction performance in regard to a specific tuning parameter, with this process of validation repeated ten times.

In order to implement the 10-fold cross validation, the training data were first partitioned into ten nonoverlapping subsamples of equal size. In the first stage, nine training subsamples were refined into balanced data through undersampling, and the model was fitted to the refined data using specific tuning parameter values. Then, the fitted model was applied to the remaining validation subsample to compute the predicted probable values for the occurrence of heavy rain damage (1), which were then compared to some specific probability cutoff values to determine the occurrence or nonoccurrence of heavy rain damage. Sensitivity and specificity were calculated in this process. At this point, by plotting sensitivity against (1 – specificity) for all possible probability cutoff values ranging from 0 to 1, we obtained a two-dimensional curve called the receiver-operating characteristic (ROC) curve, with the area under this curve being called the area under the curve (AUC).

The AUC value provides the criterion of validation. The AUC of a model can take on values from 0.5 to 1, and when its value approaches 1, the model may be judged as having superior prediction performance. AUC is widely used as a representative metric in performance comparisons because it can compare the relative prediction performance of binary classification models regardless of the probability cutoff values used in them. Moreover, by referring to the ROC curve in the validation process, the probability cutoff value that maximizes the sum of sensitivity and (1 – specificity) can be used in the developed model to classify 1 and 0 in its future predictions.

The abovementioned process was repeated ten times, and the tuning parameter values corresponding to the highest of the ten validated average AUC values were selected as optimum tuning parameter values. The whole training process was then concluded by refining the entire set of training data through undersampling and fitting the model to the data using the optimum tuning parameter

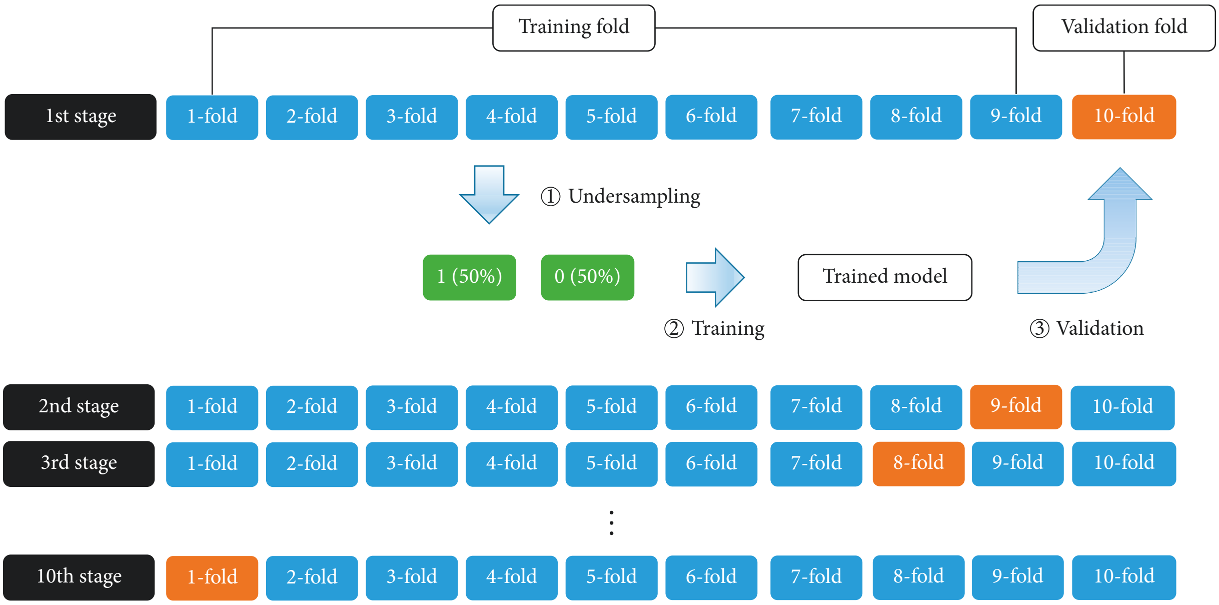


Figure 6: Concept of 10-fold cross validation.

values. Subsequently, the trained model was applied to the test data in order to evaluate its prediction performance in actual future circumstances. Figure 6 shows a conceptualization of the 10-fold cross-validation process, and Figure 7 shows a conceptual diagram of the ROC curve and the AUC.

3. Methodology

In order to develop our prediction model for heavy rain damage using machine learning based on big data, we selected the Seoul Capital Area as the study area and constructed response and explanatory variables from the data on heavy rain damage amounts given in the Annual Natural Disaster Report and the meteorological big data collected from the Korea Meteorological Administration's Open Weather Data Portal (https://data.kma.go.kr).

3.1. Selection of the Study Area. For the purpose of constructing the prediction model for heavy rain damage, our desired study area was the one that has a high incidence of heavy rain damage. We collected and analyzed the heavy rain damage data from 1994 to 2015 based on the ten regional divisions used in weather forecasts by the Korea Meteorological Administration. As shown in Table 1 and Figure 8, the Seoul Capital Area (i.e., Seoul, Incheon, and Gyeonggido) had the highest incidence of heavy rain damage, and this was chosen as the area in our study.

3.2. Constitution of the Response Variable. In order to construct the response variable of our prediction model for heavy rain damage, we collected data on heavy rain damage from 1994 to 2015 from the Annual Disaster Report provided by the Ministry of Interior and Safety (MOIS) in Korea. Data on the extent of heavy rain damage were collected by administrative region and disaster period, and they were converted into "1" on the days when heavy rain damage occurred and into "0" otherwise. Thus, response variable values of the prediction model for heavy rain damage are in the form of binary data from 1994 to 2015, and since they are drawn from daily data across a 22-year period from 66 administrative areas, they constitute a total of around 500,000 data points.

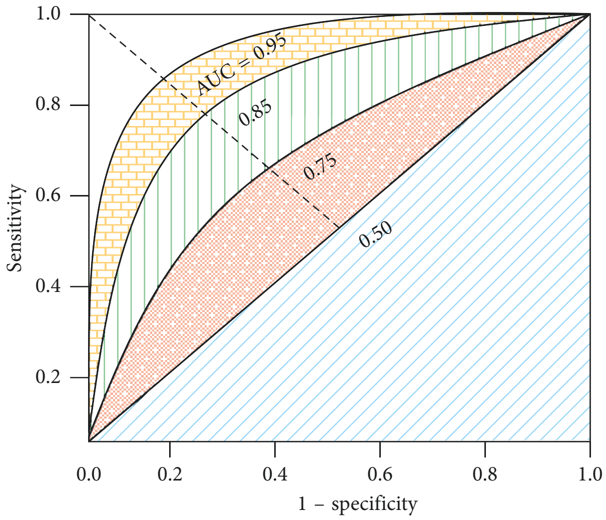


Figure 7: Concept of ROC and AUC.

3.3. Constitution of Explanatory Variables. The Korea Meteorological Administration collects large amounts of weather observation data every day from various sources such as the sky, the sea, and on the ground and produces high volumes of prediction data using over ten numerical weather forecast models. Its weather and climate data satisfy all of the four major characteristics of big data, namely, size, diversity, speed, and value, and it makes this meteorological

Table 1: Incidence of heavy rain damage.

Division	Incidence of damage
Seoul Capital Area (Seoul, Incheon, and Gyeonggido)	1,044
Gangwondo (Yeongseo)	193
Gangwondo (Yeongdong)	66
Chungcheongbukdo	210
Daejeon, Sejong, and Chungcheongnamdo	345
Jeollabukdo	250
Gwangju and Jeollanamdo	461
Daegu and Gyeongsangbukdo	343
Busan, Ulsan, and Gyeongsangnamdo	428
Jeju	34

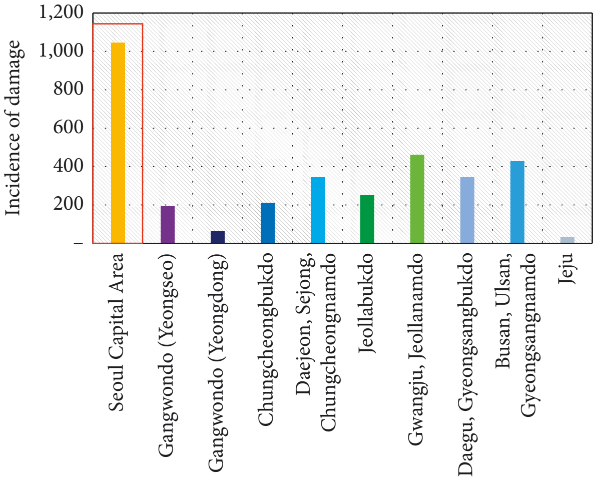


Figure 8: Incidence of heavy rain damage.

big data available for free at the Open Weather Data Portal (https://data.kma.go.kr). Accordingly, we used the Automated Synoptic Observing System's (ASOS) daily weather observation data from 1994 to 2015 available at the Open Weather Data Portal, along with data on regional characteristics as the explanatory variables (29 in total) of our prediction model for heavy rain damage. In particular, we used a total of 27 variables from the ASOS weather observation data, excluding variables that are not related to heavy rain damage, such as daily maximum fresh snow depth and daily maximum fresh snow depth time. Table 2 shows the list of explanatory variables used in this study.

3.4. Matching of Response Variable and Explanatory Variable Data. Since the ASOS weather observations used as explanatory variable data are values measured at different observatories, they need to be matched one to one with response variable data collected from cities, counties, and districts. For this purpose, we determined the boundaries of each administrative region (cities, counties, and districts), the locations of ASOS observatories, and the Thiessen areas, as shown in Figure 9. Ideally, there should just be one set of weather observation values for any one administrative division at a given period of time, but we found that there were several sets of weather observation values with the same properties for the same administrative region measured at different adjacent observatories. The Thiessen polygon method—used to obtain precipitation in catchment areas by treating the area close to an observation point as weighted (Thiessen area)—is used in the field of water resources to calculate areal precipitation, and the same methodology was used in this study. Accordingly, the observations measured for the same administrative region by different adjacent observatories were combined as a weighted average that takes into account the Thiessen area ratio. That is, observation values with the same properties for a given day were combined as a weighted average that takes into account the Thiessen area ratio. We processed the ASOS weather observation data in this way for all the cities, counties, and districts. Also, if there are missing values in the data measured by some observatories, it is not easy to modify such observation data arbitrarily, nor is it easy to verify the modified data. Therefore, in such cases, we only considered the weights (Thiessen area ratio) for data from observatories without missing values, readjusting their values to sum up to 1 before calculating the weighted average. However, if the data from all observatories had missing values, then we treated the corresponding sample (row of observations) as missing since the weighted average could not be obtained in such cases. A total of 530,317 samples were obtained by matching the response variable data with the explanatory variable data. Among these, there were 1,796 samples with missing values, but we found that none of them had a response variable value of 1 (occurrence of heavy rain damage). Since samples that have missing data from all observatories ended up being deleted in the process of constructing the prediction model, we removed the samples with missing data in order to facilitate the analysis and used a total of 528,521 samples for the analysis.

3.5. Types of Data. The types of data used in the analysis are shown in Table 3. The response variable has only two values, 1 (occurrence of heavy rain damage) and 0 (no heavy rain damage), and there were a total of 6,651 cases with value 1, corresponding to about 1.3% of the total number of cases.

First, for the purpose of developing the model, we used data from 1994 to 2011 as training data and data from 2012 to 2015 as test data. We have selected data from the year 2012 as the test data since there were many occurrences of heavy rain damage that year, thus allowing us to impose a stricter test of the prediction performance of models. The ratios between values 0 and 1 for the training data and the test data are shown in Table 4.

As shown in Table 4, there is a great imbalance in the ratios between values 1 and 0 for both the training and the test data. As explained in Section 2.5, the model is developed using the training data, and then its prediction performance is evaluated in terms of the AUC value using the test data.

3.6. Definition of Prediction Models. The kinds of data used in this study are the daily data on whether there was heavy rain damage and the corresponding weather observation

Table 2: List of explanatory variables.

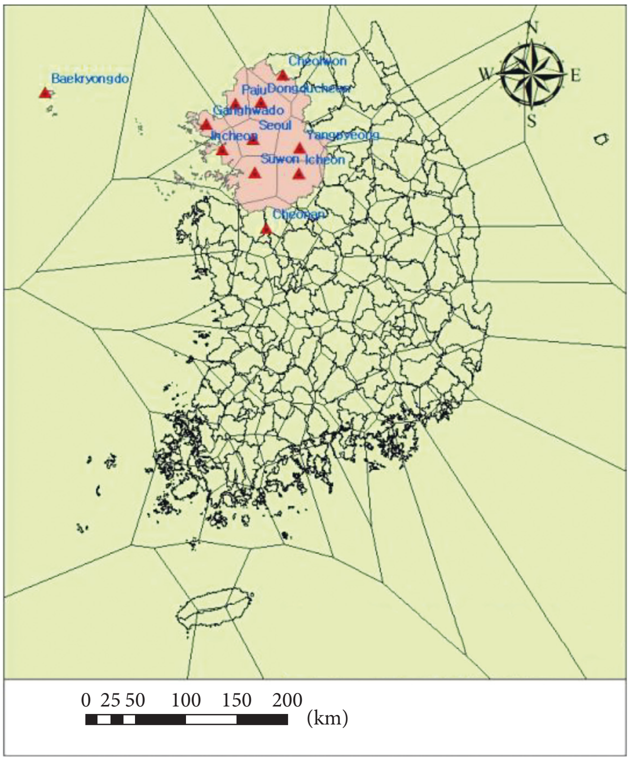
Category	Variables
Temperatures	Average temperature (°C) Minimum temperature (°C) Maximum temperature (°C)
Precipitation	Precipitation duration (hr) 10-minute maximum precipitation (mm) 1-hour maximum precipitation (mm) Daily precipitation (mm)
Humidity	Average dew point temperature (°C) Minimum relative humidity (%) Average relative humidity (%) Average vapor pressure (hPa)
Insolation and insolation duration	Possible duration of sunshine (hr) Duration of sunshine (hr)
Fog	Fog duration time (hr)
Evaporation	Large evaporation (mm) Small evaporation (mm) 9-9 precipitation (mm)
Wind	Maximum instantaneous wind speed (m/s) Maximum wind speed (m/s) Average wind speed (m/s) Wind match (100 m)
Atmospheric pressure	Average local pressure (hPa) Maximum sea-level pressure (hPa) Minimum sea-level pressure (hPa) Average sea-level pressure (hPa)
Cloud	Average total cloud amount (1/10) Average middle and lower layers cloud amount (1/10)
Regional characteristics	Administrative area (km^2) Area classification category variable

data. The algorithms for developing a model from these data can be sorted into two kinds according to the kind of weather observation data they use. First, let us stipulate that the algorithm using the weather observation data on a given day to predict heavy rain damage on that same day is Algorithm 1. Let the actual occurrence or nonoccurrence of heavy rain damage on a given day be y_t, let that same day's weather observation data be the vector $x_t = (x_{1t}, x_{2t}, \ldots, x_{pt})$, and let the predicted occurrence or nonoccurrence of heavy rain damage for the same day, computed by entering that day's weather observation data into model $f(\cdot)$, be $\hat{y}_t$. Then, Algorithm 1 can be expressed as

$$\hat{y}_t = f(x_t). \tag{3}$$

Using weather observation data on a given day to predict whether or not there will be heavy rain damage that day is the ideal case of prediction, and it is expected to have a high prediction performance. However, even if the prediction performance of the model based on Algorithm 1 is evaluated highly, it will be almost impossible in practice to use weather observation data on a given day in the same-day forecasting of heavy rain damage due to the physical time difference. The practical problem is that the weather observation data to be used in the same-day prediction cannot be observed because it lies in the future from the standpoint of analysis. To overcome this problem, we can consider using weather prediction data that predict the weather for the target area, but the weather forecast data being provided in Korea are very limited in number and are known to have very high uncertainty. Figure 10 conceptualizes Algorithm 1, the ideal kind of prediction model.

Therefore, we need another algorithm that is more realistic and whose prediction performance is not much lower than that of Algorithm 1. For this purpose, the present study also considered Algorithm 2, which can be used to develop a model that uses past observation data (from one to k days



▲ ASOS in the metropolitan area
🟥 Thiessen area of the metropolitan area
🟩 Thiessen area of the total area

Figure 9: Administrative boundary and ASOS location.

Table 3: Types of data for analysis.

Division	Explanation	Period	Remarks
Response variable	Whether heavy rain damage occurred	1994–2015	Binary data of 1 and 0 1: 6,651 0: 521,870
Explanatory variables	Weather observation data	1994–2015	Continuous data

Table 4: Types of training data and test data.

Division	0 (no damage)	1 (heavy rain damage)
Training data	426,654 (98.62%)	5,987 (1.38%)
Test data	95,216 (99.30%)	664 (0.70%)

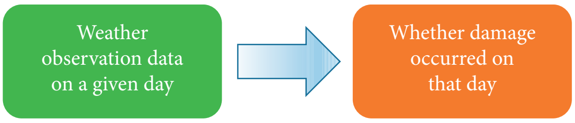

Weather observation data on a given day
Whether damage occurred on that day

Figure 10: Algorithm 1 concept.

ago) to predict heavy rain damage on a given day, and this algorithm can be expressed as

$$\hat{y}_t = f(x_{t-1}, x_{t-2}, x_{t-3}, \ldots, x_{t-k}). \tag{4}$$

The reason for developing the prediction model using Algorithm 2 is that past weather observation data are relatively highly reliable and offer a rich variety of usable data. Therefore, if we develop a prediction model based on Algorithm 2 that has almost the same prediction performance as a model based on Algorithm 1, then Algorithm 2 can be a good alternative to Algorithm 1. In the process of developing such a model, we also performed an analysis on how many days ago we have to look back to in the use of past weather observation data for optimal prediction performance. Figure 11 conceptualizes Algorithm 2 as a realistic kind of prediction model.

4. Results and Discussion

We developed the prediction model for heavy rain damage based on big data by constructing machine learning models for each of the two algorithms. The prediction performance of the models was evaluated by computing the AUC value for each algorithm and model, and the final model was selected on this basis. We checked whether this final model's AUC value is maintained at a constant level by examining the variability in results that can occur during the sampling process.

4.1. Development of Prediction Models for Heavy Rain Damage. The models based on decision tree learning and related ensemble techniques such as bagging, random forests, and boosting were trained on the training data. The models used in the training all had their own tuning parameters. The optimum tuning parameter values that must be trained for each model are as follows: decision trees must have optimal depth for pruning, random forests must have an optimal number of explanatory variables, and boosting must have optimal depth and learning rate. In order to develop a model with high prediction performance, there needs to be appropriate training in these tuning parameters, and the 10-fold cross-validation method described in Sections 2.2 through 2.5 was used to train the optimum tuning parameter values. Next, the model was fitted to the entire training dataset using the optimum tuning parameter values and then applied to the test data to evaluate its prediction performance.

4.2. Evaluation of Prediction Performance. In order to evaluate the prediction performance of a model in regard to the binary data of 1 (occurrence of heavy rain damage) and 0 (no heavy rain damage), the model developed using the training data was applied to the test data to compute its AUC value. The model's AUC value in regard to the test data affords an evaluation of the model's prediction performance in relation to actual future situations and provides a measure for comparing the real-world performance of different models.

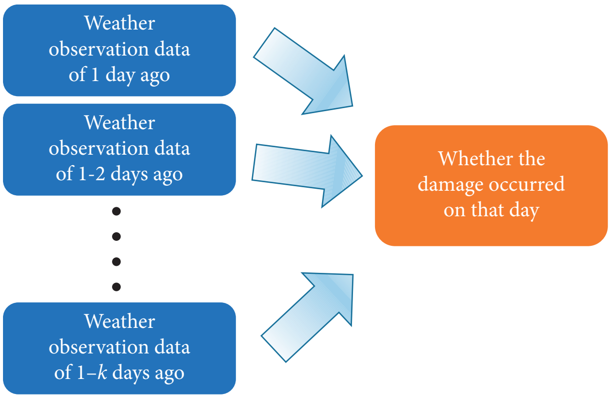

Weather observation data of 1 day ago
Weather observation data of 1-2 days ago
Weather observation data of 1–k days ago
Whether the damage occurred on that day

Figure 11: Algorithm 2 concept.

In this study, we developed prediction models based on Algorithm 1 and Algorithm 2, and each of these models was applied to the test data to compute its AUC value. Now, if the AUC value of an Algorithm 2 model is equal to that of an Algorithm 1 model, then it is possible to conclude that the Algorithm 2 methodology proposed in this study can be used in the development of an actual prediction model. Figure 12 is a conceptualization of the process of comparing the two kinds of models through evaluation of their prediction performance.

4.3. Selection of the Final Model through Evaluation of Prediction Performance. The decision tree, bagging, random forest, and boosting models based on Algorithm 1 and Algorithm 2 were trained on the training data using 10-fold cross validation. Then, the trained models were applied to the test data to yield their AUC values, as shown in Table 5.

On the whole, it can be seen that the prediction models based on Algorithm 1 that predict heavy rain damage using same-day weather data have high AUC values. However, as described in Section 3.6, Algorithm 1 is not usable in practice. Consequently, we need to find a model based on Algorithm 2 with an AUC value closest to that of the Algorithm 1 models. Among the evaluations of the prediction performance of Algorithm 2 models, it was shown that the boosting model fitted to weather observation data of 1 to 4 days ago has the highest AUC value (95.867%). In the case of the random forest model based on Algorithm 2, it is suspected that the model is overfitting the training data because its AUC value gets lower as it uses more past data, and it is shown in all the cases that the random forest model has a lower prediction performance than the boosting model. Therefore, the present study selected the boosting model fitted to weather observation data of 1 to 4 days ago as the final prediction model. Although it does not have as high an

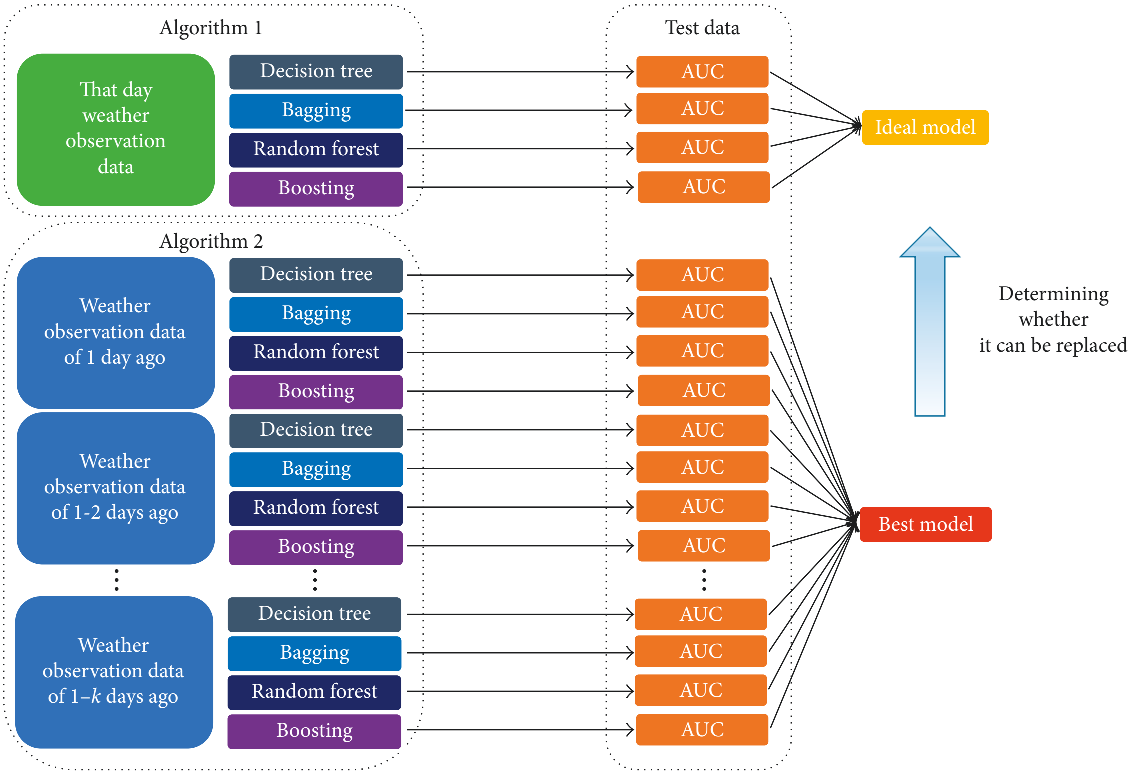

Algorithm 1
That day weather observation data
Decision tree
Bagging
Random forest
Boosting
Algorithm 2
Weather observation data of 1 day ago
Weather observation data of 1-2 days ago
Weather observation data of 1–k days ago
Test data
AUC
Ideal model
Best model
Determining whether it can be replaced

Figure 12: Predictive evaluation concept.

Table 5: Evaluation of prediction performance by model.

Division	Time	Decision tree (%)	Bagging (%)	Random forest (%)	Boosting (%)
Algorithm 1	That day	94.072	96.371	96.518	96.345
Algorithm 2	1 day ago	90.923	95.048	95.028	95.408
	1-2 days ago	90.914	94.775	94.971	95.646
	1–3 days ago	90.357	94.420	94.698	95.516
	1–4 days ago	90.438	94.632	94.666	**95.867**
	1–5 days ago	90.154	94.678	94.445	95.772
	1–6 days ago	90.120	94.233	94.244	95.511
	1–7 days ago	90.719	92.915	93.117	94.914

AUC value as those of the Algorithm 1 models, it can replace these models in the sense that, generally speaking, an AUC value above 95% guarantees the prediction performance of a model.

4.4. Validation of Undersampling. In this study, models are developed by reducing the size of the data through undersampling during the training process, and then they are applied to the test data to evaluate their prediction performance. Since there is the possibility that a model with high prediction performance may have been developed by chance during the sampling process, the prediction performance of the final model was evaluated several times in order to examine the variability of results that could occur during the sampling process and to check whether the model's AUC value is maintained at a constant level. For this purpose, we trained the final model that is, the boosting model fitted to weather observation data of 1 to 4 days ago, 20 times in the same way. Since random undersampling is applied during the learning process, the 20 resulting models will differ in their prediction performance. It is possible to determine whether the prediction performance is maintained at a constant level by using the test data on the 20 models to compute their AUC values. As shown in Table 6, the mean AUC value was 95.55%, and the standard deviation was 0.25%, thus indicating that there was not much variability. Therefore, although the final model in this study was developed not by using the entire training dataset but through undersampling, we may conclude that it is a superior model with a high prediction performance that is guaranteed.

4.5. Summary. The main results of this study can be summarized as follows:

(1) To develop our prediction model for heavy rain damage using machine learning based on big data,

Table 6: Evaluation of repeat prediction performance of final model.

Division	Time	Average AUC	Standard deviation of AUC
Algorithm 2	1–4 days ago	95.55%	0.25%

we selected the Seoul Capital Area as the study area and constructed the response and explanatory variables by collecting relevant data from the Annual Natural Disaster Report provided by the Ministry of the Interior and Safety and the meteorological big data provided at the Open Weather Data Portal.

(2) Two algorithms were derived, namely, Algorithm 1 that predicts heavy rain damage on a given day using same-day weather observation data and Algorithm 2 that predicts heavy rain damage on a given day using past weather observation data. Each of these algorithms was used in machine learning (decision tree, bagging, random forest, and boosting) to develop a prediction model for heavy rain damage.

(3) Evaluation of the prediction performance of each algorithm and model showed that most of these models have high prediction performance with AUC values greater than 90%. Algorithm 1 models had the highest AUC values, but they were not significantly different from the AUC values of Algorithm 2 models.

(4) Therefore, it was determined that Algorithm 2 can substitute for Algorithm 1, and *the boosting model using past weather data from one to four days ago*, which had the highest prediction performance among Algorithm 2 models, *was selected as the final model*.

(5) The final model maintained its average AUC value of 95.55% at a constant level in a test of its variability in repeated sampling, and thus it was deemed to be a superior model with guaranteed prediction performance.

5. Conclusion

We developed a model for the prediction of heavy rain damage based on the big data provided by the Korea Meteorological Administration and machine learning that can maximize the prediction performance of the model, and the model could be used in implementing a proactive disaster management system. However, this study has some limitations on the number of damage data and the use of hydrometeorological data. We used heavy rain damage data of 22 years from 1994 to 2015, which were provided in Annual Natural Disaster Report published by the Korean government, as the dependent variables of the model. Actually, there are more damage data but the reliable data were used in this study. Therefore, if we have more data, we can get better prediction performance of the model. Also this study just used hydrometeorological data such as temperature, precipitation, humidity, and so on as the independent variables of the model. Therefore, we may need more data which are related to disaster prevention projects, disaster recovery budget, and socioeconomic factors such as increasing ratio of impervious area, per capita income, and ratio of vulnerable populations of local governments. Taking into account such damage-related factors in addition to hydrometeorological factors will make it possible to develop a more reliable prediction model for heavy rain damage.

Previous studies have mostly considered just one to three independent or explanatory variables and have used only linear methods such as linear regression analysis. The present study, however, applies the technique of machine learning along with big data for development of heavy rain damage model, and we believe that the results of this study represent a tangible advance in this area. The prediction model for heavy rain damage presented in this study can be utilized to provide a heavy rain damage prediction service without much additional cost and the occurrence probability of heavy rain damage for each administrative region or local governments.

The proposed service comprises three main stages. First, it collects weather observation data of one to four days before predicting heavy rain damage occurrence. Then, the collected data are entered into the prediction model developed for heavy rain damage prediction. In the final stage, the prediction is made for local governments. Say, if the developed model predicts the heavy rain damage occurrence in a city, county, or district, then the specified region will be able to take preparation for disaster and disaster management measures in advance (inspection of vulnerable areas and facilities, placement of emergency personnel, announcement of emergency evacuation instructions, and so on), thus resulting in a significant reduction of heavy rain damage.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Inha University Research Grant.

References

[1] MOIS (Ministry of the Interior and Safety), *The 2015 Annual Natural Disaster Report*, MOIS, Republic of Korea, 2016.

[2] S. A. Davis and L. L. Skaggs, *Catalog of Residential Depth-Damage Functions Used by the Army Corps of Engineers in Flood Damage Estimation*, Army Engineer Institute for Water Resources, Alexandria, VA, USA, 1992.

[3] B. G. Chae, W. Y. Kim, Y. C. Cho, K. S. Kim, C. O. Lee, and Y. S. Choi, "Development of a logistic regression model for

probabilistic prediction of debris flow," *Journal of Engineering Geology*, vol. 14, no. 2, pp. 211–222, 2004.

[4] B. J. Kim, J. H. Song, H. S. Kim, and I. P. Hong, "Parameter calibration of storage function model and flood forecasting: (2) comparative study on the flood forecasting methods," *Journal of the Korean Society of Civil Engineers B*, vol. 26, no. 1, pp. 39–50, 2006.

[5] R. J. Murnane and J. B. Elsner, "Maximum wind speeds and US hurricane losses," *Geophysical Research Letters*, vol. 39, no. 16, 2012.

[6] B. F. Prahl, D. Rybski, J. P. Kropp, O. Burghoff, and H. Held, "Applying stochastic small-scale damage functions to German winter storms," *Geophysical Research Letters*, vol. 39, no. 6, 2012.

[7] A. R. Zhai and J. H. Jiang, "Dependence of US hurricane economic loss on maximum wind speed and storm size," *Environmental Research Letters*, vol. 9, no. 6, article 064019, 2014.

[8] J. S. Lee, G. Eo, C. H. Choi, J. W. Jung, and H. S. Kim, "Development of rainfall-flood damage estimation function using nonlinear regression equation," *Journal of the Korean Society of Disaster Information*, vol. 12, no. 1, pp. 74–88, 2016.

[9] J. S. Kim, C. H. Choi, J. S. Lee, and H. S. Kim, "Damage prediction using heavy rain risk assessment: (2) development of heavy rain damage prediction function," *Journal of Korean Society of Hazard Mitigation*, vol. 17, no. 2, pp. 371–379, 2017.

[10] C. H. Choi, J. S. Kim, J. H. Kim, H. Y. Kim, W. J. Lee, and H. S. Kim, "Development of heavy rain damage prediction function using statistical methodology," *Journal of Korean Society of Hazard Mitigation*, vol. 17, no. 3, pp. 604–612, 2017.

[11] T. H. Choo, K. S. Kwak, S. H. Ahn, D. U. Yang, and J. K. Son, "Development for the function of wind wave damage estimation at the western coastal zone based on disaster statistics," *Journal of the Korea Academia-Industrial Cooperation Society*, vol. 18, no. 2, pp. 14–22, 2017.

[12] C. Dorland, R. S. Tol, and J. P. Palutikof, "Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change," *Climatic Change*, vol. 43, no. 3, pp. 513–535, 1999.

[13] R. A. Pielke Jr. and M. W. Downton, "Precipitation and damaging floods: trends in the United States, 1932–97," *Journal of Climate*, vol. 13, no. 20, pp. 3625–3637, 2000.

[14] H. Toya and M. Skidmore, "Economic development and the impacts of natural disasters," *Economics Letters*, vol. 94, no. 1, pp. 20–25, 2007.

[15] R. Mendelsohn and G. Saher, *The Global Impact of Climate Change on Extreme Events*, World Bank, Washington, DC, USA, 2011.

[16] J. Liu, "Weather or wealth: an analysis of property loss caused by flooding in the US," in *Proceedings of the 2012 Annual Meeting on Agricultural and Applied Economics Association*, Seattle, WA, USA, August 2012.

[17] G. Furquim, G. Pessin, B. S. Fai, E. M. Mendiondo, and J. Ueyama, "Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory," *Neural Computing and Applications*, vol. 27, no. 5, pp. 1129–1141, 2016.

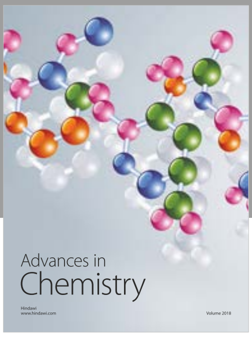
[18] K. M. Asim, M. Mart, F. MartMar, A. Basit, and T. Iqbal, "Earthquake magnitude prediction in Hindukush region using machine learning techniques," *Natural Hazards*, vol. 85, no. 1, pp. 471–486, 2017.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

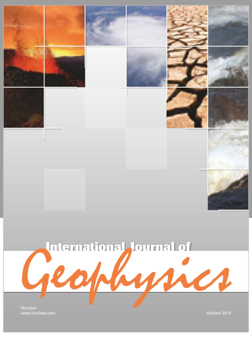
[20] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the International Conference on Machine Learning*, vol. 96, pp. 148–156, Atlanta, GA, USA, June 1996.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

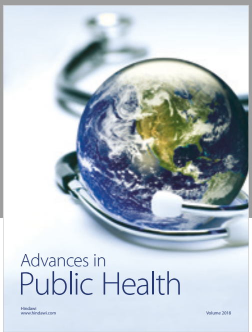
[22] R. Longadge, S. Dongre, and L. Malik, "Class imbalance problem in data mining: review," *International Journal of Computer Science and Network*, vol. 2, no. 1, pp. 1–6, 2013.

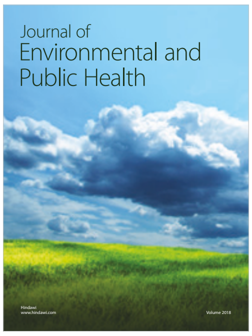

Advances in
Chemistry

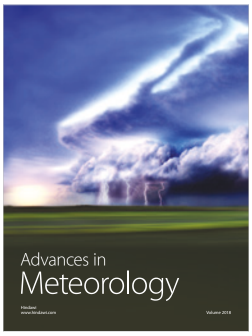
Scientifica

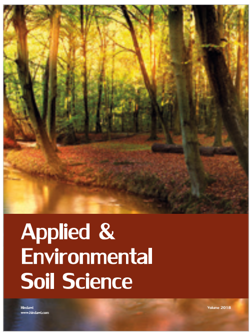

International Journal of
Geophysics

The Scientific
World Journal


Advances in
Public Health

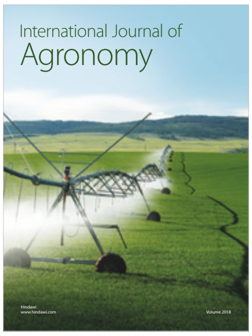

Journal of
Environmental and
Public Health


Advances in
Meteorology


Applied &
Environmental
Soil Science

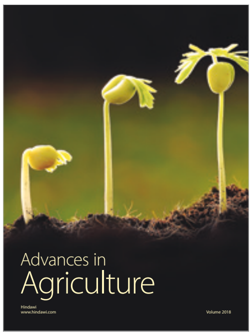
Journal of
Geological Research

International Journal of
Forestry Research

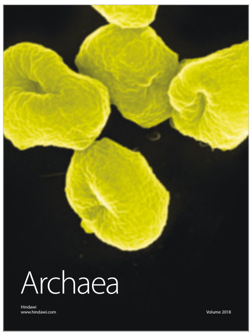

International Journal of
Agronomy

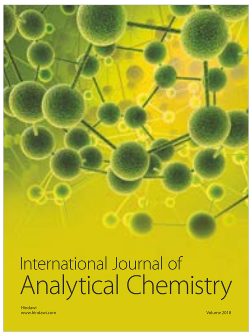
International Journal of
Microbiology

International Journal of
Ecology


Advances in
Agriculture

International Journal of
Biodiversity


Archaea


International Journal of
Analytical Chemistry

Journal of
Chemistry

Journal of
Marine Biology