Sensors and Materials, Vol. 35, No. # (2023) 1–##
MYU Tokyo

S & M ####

Fire Risk Prediction Analysis Using Machine Learning Techniques

Min Song Seo[1], Ever Enrique Castillo-Osorio[2], and Hwan Hee Yoo[1*]

[1]Department of Urban Engineering, Gyeongsang National University,
501, Jinju daero, Jinju, Gyeongsangnam-do, 660701, Korea
[2]School of Civil & Environmental Engineering, Yonsei University,
50, Yonsei-ro, Seodaemungu, Seoul 03722, Korea

(Received November 16, 2022; accepted March 16, 2023; online published April 10, 2023)

Keywords: fire property damage, support vector machine, random forest, gradient-boosted regression tree, k-fold cross-validation

The damage caused by fire accidents is increasing worldwide. In particular, when a fire occurs, property damage directly affects the lives of citizens. Therefore, in this study, machine learning techniques were applied to analyze the prediction of the future amount of property damage from fire as well as the fire occurrence factors within a geographic area. To achieve this, three years of spatially distributed fire big data for Seoul, the capital of Korea, was used. For the predictive analysis of the amount of fire property damage, the results of analysis by applying machine learning techniques through k-fold cross-validation were calculated. As part of these results, when predicting and analyzing the amount of fire property damage using the random forest (RF) algorithm, an accuracy of 83% was calculated by comparing the predicted data with the actual data. On this basis, the importance of the fire risk factors was analyzed, and it was found that the main factor in the occurrence of fires is the condition of the facilities inside apartment houses. The findings of this study are expected to be used as an important guide for identifying property damage by fire and the factors determining the occurrence of fires in Korea, enabling the evaluation of their spatial distribution and the application of corrective measures to reduce possible damage by urban fires.

1. Introduction

Fire is a hazard in all countries and causes a large amount of damage every year. Fire accidents increase social anxiety and threaten the lives and property of citizens. Although fire accidents vary according to topographical, geographical, and cultural factors, many casualties and extensive property damage occur worldwide. In addition to the property damage caused by fires, indirect costs include disruption to business, reduced creditworthiness, fire site clearance costs, and human loss.

The National Fire Administration of Korea and the Fire and Disaster Management Agency of Japan under the Ministry of Internal Affairs and Communications collect and compile data on

*Corresponding author (BK21+, ERI): e-mail: hhyoo@gnu.ac.kr

ISSN 0914-4935

all fire accidents. In the USA, the National Fire Protection Association (NFPA) and the US Fire Administration (USFA) collect fire data through the NFPA Fire Survey program. The fire property damage statistics in Korea are not very specific, imposing limitations when accurately analyzing property damage caused by fires. In addition, property damage from fire as a proportion of gross domestic product (GDP) is declining in the USA and Japan but increasing in Korea, and since 1994, fire damage in Japan and that in the USA have accounted for similar proportions of GDP. Additionally, after taking inflation (based on the consumer price index) into account, property damage due to fires is decreasing in Japan and the USA but increasing in Korea.[1]

In Korea, fires are one of the most frequent social disasters along with traffic accidents. In the event of a fire accident, it is important to determine the initial cause by carefully analyzing the fire data. From 2017 to 2021, the total number of accidents was 201,545, the number of casualties (injuries and deaths) was 11,718, and the total value of property damage was $2.7 billion (Table 1).

In Korea, the number of fire accidents has recently decreased each year. However, the value of property damage has increased by 116% in the last five years.[2] Considerable human and material damage occurs in Korea owing to the many fire accidents every year.[2] In particular, in Seoul, the capital of Korea, there were 5978 fire accidents causing $12.21 million of damage in 2017, 6368 accidents causing $16.92 million of damage in 2018, and 5881 fire accidents causing $73.77 million of damage in 2019; although the number of fire accidents has fluctuated, the value of property damage was 496% higher in 2019 than in 2017. Therefore, various countermeasures are needed to reduce this trend. According to the official fire situation report of the NFPA, there were approximately 1338500 fire accidents in the USA in 2020, resulting in 3500 deaths, 15200 injuries, and direct property losses of $21 billion, with about 74% of fires occurring in urban areas.[3]

According to the above statistics, the risk of fire accidents occurring in urban areas is increasing worldwide. Therefore, research on fire accidents is required to reduce the damage caused by them. Recently, various studies on fires have been conducted, such as fire damage, fire risk, fire risk area spatial, fire statistical, and fire prediction analyses. In this study, to identify research trends based on these techniques, we reviewed previous studies, specifically those on the statistical and predictive analyses of fire outbreaks. First, we reviewed studies on the statistical analysis of fire occurrence in Korea. Yun *et al.* presented a disaster risk assessment model limited to fire, facility, and evacuation risks for Cheongju City in Korea as an example. They evaluated the fire risk in every small urban sector, called a *dong* in Korea, on the basis of the number of fire occurrences and the amount of damage. This fire risk was found to be relatively high in urbanized areas and the central commercial area with a high density of low-

Table 1
Recent statistics of fire accidents in Korea.

	2017	2018	2019	2020	2021	Total
Number of fires	44178	42338	40103	38659	36267	201545
Number of casualties	2197	2594	2515	2282	2130	11718
Property damage	$385189.97	$425326.75	$652348.02	$456284.19	$835197.57	$2754361.70

income residential areas and public facilities.[4] Kang found a high correlation (0.97) between population and fire occurrence by analyzing the Fire Prevention District Current Status and Improvement Plan. These results allowed the author to predict the number of fire occurrences in the study area.[5] In addition, we reviewed papers examining fire incidence in other countries. Chhetri *et al.* used multiple regression analysis and the analysis of variance (ANOVA) to study the relationship between socioeconomic characteristics, such as income, education, occupation, welfare, living conditions, and so forth, and fire incidence rates in southeast Queensland, Australia. On the basis of their results, they presented a model that could predict the fire rate from the number of unemployed people, the number of indigenous people, the number of single-parent families with children, and the proportion of families living in segregated dwellings.[6] Hastie and Searle used principal component analysis (PCA) and ordinary least squares (OLS) regression to analyze fire incidence according to socioeconomic factors (poverty, low housing quality, unemployment rate, low education level, ethnicity) in residential areas in the West Midlands, UK. They revealed that the racial composition and unemployment rate in the West Midlands have a significant impact on fire rate.[7] Jennings reviewed the literature related to the analysis of social and economic characteristics of residential fire risk in urban areas. He commented on studies analyzing the socioeconomic characteristics that affect fire occurrence and damage, and suggested the establishment of a fire prevention plan and the direction of future research to apply the results of these studies.[8] In addition, we reviewed studies on prediction analysis according to fire occurrence. Lee *et al.* employed the amount of property damage and fire occurrence in Seoul to derive the fire risk by building. They utilized variables such as the characteristics of the administrative district to which the building belongs and accessibility to firefighting facilities as well as the characteristics of the building. As a result of using the random forest (RF) algorithm, they obtained an accuracy of about 74%, and the variable of building characteristics was found to be important. Using the constructed model, they predicted the fire risk for 300 buildings in Seoul.[9] Madaio *et al.* presented a model that predicts the number of fires in Atlanta, USA, by using the support vector machine (SVM) and RF machine learning algorithms. The variables included the area and number of floors of buildings. The results of their analysis can be used to measure the fire risk of buildings and to prioritize fire inspections.[10]

As outlined above, many studies related to fire have been conducted and numerous studies are still in progress. Most of the reviewed papers involving statistical analysis according to fire occurrence identified the relationship between fire occurrence and various factors based on a simple statistical analysis, and in the case of predictive analysis, the degree of fit was calculated for each model. In previous research on fire occurrence prediction, studies on the development of fire prediction models using various machine learning algorithms have been conducted, but a limitation of these studies is that an accurate relationship between the factors determining fire occurrence and fire damage cannot be identified. Therefore, in this study, to overcome this limitation, we collected spatially distributed fire damage big data in Seoul and adopted fire occurrence factors without multicollinearity through statistical techniques such as correlation and multiple regression analyses. As machine learning techniques, we applied and analyzed SVM, which has high accuracy without overfitting, RF, which is not sensitive to noise and

outliers, and gradient-boosted regression tree (GBRT), an ensemble method that is newer and more advanced than artificial neural networks or decision trees. Moreover, we performed k-fold cross-validation to increase the accuracy of machine learning, and we conducted the predictive analysis of fire occurrence factors, focusing on the adopted technique. In addition, when a fire accident occurs, we predict the amount of fire damage with the greatest impact and the spatial area where the greatest damage occurs, as well as the main factor affecting the amount of fire damage by targeting the risk area. Our aim is not only to reduce the frequency of fires, but also to assess their spatial distribution and apply corrective measures to reduce fire damage. A flowchart of the processes considered in this study is shown in Fig. 1.

2. Methodology

2.1 Evaluation of predictive power using machine learning techniques

The sufficiency of the input training data and the abundance of predictors are greatly affected by the hyperparameters of the training model. Therefore, a hyperparameter optimization process is required.[11,12] There are no established standards or methods for selecting hyperparameters in machine learning. It is common to find the hyperparameter that minimizes the error by changing it according to the actual data used in machine learning.[11] The optimal hyperparameter can be derived through various trial and error tests. We used k-fold cross-validation[13] to select the best model and collect thresholds and weights as initial values, greatly improving the error correction. The verification process was repeated k times, and the hyperparameter with the lowest generalization error was determined as the final model. We divided the total fire occurrence data into training and test data in the ratios of 7:3 and 8:2. In the case of 7:3, 70% was analyzed as training data and the remaining 30% as test data. In addition, in the case of 8:2, 80% was

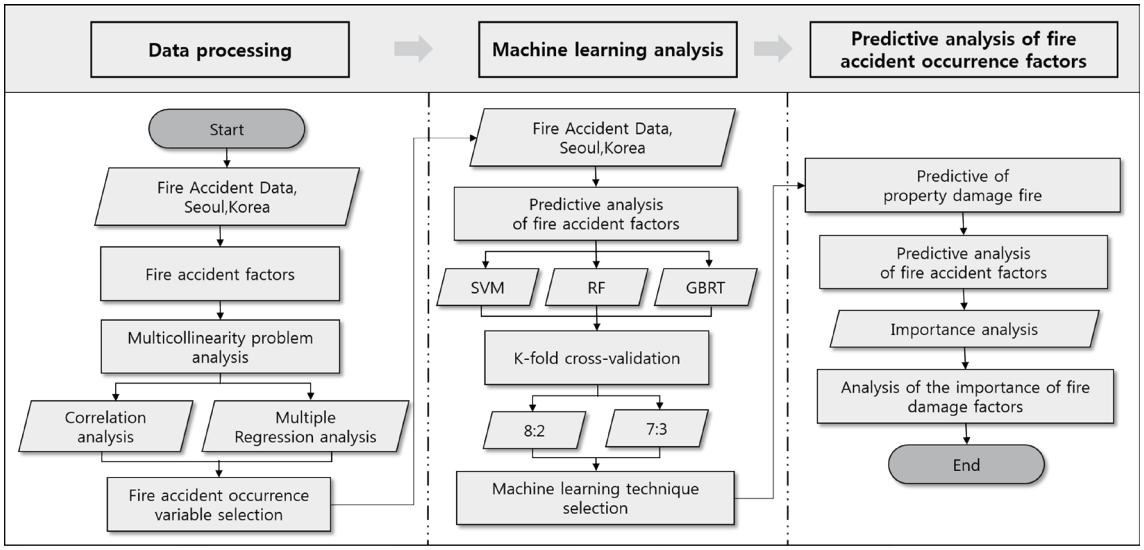


Fig. 1. Flowchart of the study.

analyzed as training data and the remaining 20% as test data. The analysis was conducted through the R Studio program, focusing on the model with the highest significance. In addition, since the accuracy of cost prediction must be compared and analyzed among models under the same conditions, we used the same training and test data for all models. In algorithm searches and modeling in fire prediction analysis using machine learning, an algorithm for practical operation is selected and a decision is made on how to model the algorithm. The selection criterion for the algorithm and modeling method must be "applicable". Therefore, it is necessary to review how to judge "availability". To use machine learning in predicting fire risk, the accuracy of the machine learning model must be recognized. The mean average error (MAE) and root mean square error (RMSE) for each model are often used to evaluate the accuracy of numerical prediction models. They can also be used to evaluate the accuracy of machine learning models. RMSE is one of the widely applied error index statistics.[14] It is generally accepted that a lower RMSE indicates a higher model efficiency. What is considered a low RMSE is defined on the basis of the standard deviation of observations.[15] The MAE is another error metric often used in model evaluation, where a value of zero indicates a perfect fit. The lower the RMSE and MAE values of the calculated data, the better the model evaluation.[16]

In this study, we performed 10-fold cross-validation while changing the hyperparameters, we set the search range of the hyperparameters, we applied those with the lowest MAE and RMSE of the validation data to the test data, and we determined the hyperparameters for each model. However, because the MAE and RMSE do not change markedly with the hyperparameters and overfitting may occur, we selected a preliminary model in addition to the model in which the MAE and RMSE are minimized. Accordingly, for the test data in this study, we selected two models for each machine learning method.

2.2 SVM

SVM is a technique for mapping and classifying vectors that are difficult to separate in a low-dimensional space (input space) or have a nonlinear distribution into a high-dimensional space (feature space). SVM is a nonlinear generalization algorithm developed by Vapnik and Lerner in 1963 and Vapnik and Chervonekis in 1964 and is a solid foundation for statistical learning theory.[17] The SVM linear regression algorithm is expressed in two equations. Equation (1) shows x in the form of a linear function. Optimizing w is equivalent to minimizing Eq. (2).[18]

$$f(x) = \langle w, x \rangle + b \tag{1}$$

$$\|w\|^2 = \langle w, w \rangle \tag{2}$$

In SVM, because low-dimensional data are mapped to high-dimensional data values, problems such as an increase in computational complexity may occur and are solved by using a kernel function. Kernel functions include a linear kernel, a sigmoid kernel, a polynomial kernel, and a Gaussian radial basis function kernel. Since there are no reasonable rules about which

kernel function to use among them and there is no significant difference in their performance, the decision is based on various tests and evaluations considering the shape of the data, the total amount of training data, and the relationship between attributes. The basic SVM is widely used in binary classification problems, and it divides one side into a positive class and the other side into a negative class around a hyperplane.[18] The most basic idea of SVM is that data consist of two classes (positive and negative classes). The goal is to find the hyperplane that best separates them. SVMs, when used for classes, aim to find the hyperplane with the maximum distance to the nearest sample point.[19]

2.3 RF

RF was proposed by Breiman in 2001 and uses the bootstrap method as an ensemble learning model.[20] A decision tree model composed of one tree has the advantage of being easy to understand because the concept is simple and visualization is possible. However, since the variance of the RF forecast is high, the stability and accuracy of the estimated forecast result are lower than those of other nonlinear models. To overcome this limitation, the results of multiple trees instead of a single tree are synthesized to increase the accuracy. RF is a combination of multiple decision tree models and creates decision trees using bootstrapped training data in the same way as bagging.[21] To solve the problem of the high variance of the data when single decision tree models are used, after securing B data sets through bootstrapping, calculating the results of each B regression tree, and finally averaging them, the final predicted value is determined. This method is called bagging. In Eq. (3). f represents the prediction model, B represents the total number of learning data, and b represents each dataset.[20]

$$\widehat{f_{bag}(x)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{f_b}(x) \tag{3}$$

An RF model that combines multiple trees is more stable than the decision tree model, and the predictive power of the model is significantly improved. In addition, overfitting can be avoided by the law of large numbers, and such an RF model is not greatly affected by noise or outliers.[20]

2.4 GBRT

Boosting is a method of making a more accurate and stronger learner by grouping relatively inaccurate weak learners. This method also combines several decision trees, similarly to bagging or RF, but unlike these methods, which create individual trees by bootstrapping, boosting continuously grows trees while modifying the original data. For example, a weakness in the first tree model is compensated for using a subsequently created tree, and a final tree model is created and weaknesses in the existing tree are supplemented or corrected by updating the residuals. AdaBoost, gradient boosting, and stochastic gradient boosting are the most frequently used

boosting algorithms. GBRT is an ensemble method based on the classification and regression tree (CART) algorithm.[22] GBRT combines two techniques: boosting and regression. Combining these techniques improves the accuracy of the model and reduces the variance.[23] Similarly to other boosting methods, GBRT trains multiple CART base learners over multiple iterations and finally generates a strong learner from a linear combination of these weak learners. GBRT, similarly to RF, can be applied to both classification and regression, has high accuracy, and does not require scaling. It also works well for continuous characteristics. Equations (4)–(7) describe a gradient boosting machine algorithm proposed by Friedman.[24] Equation (4) is the initial model consisting of only constant terms, r is an explanatory variable, x is a dependent variable, and y is a differentiable loss function. The pseudo-residuals are calculated M times using Eq. (5), r_m is calculated by fitting Eq. (6), and the residuals are updated as shown in Eq. (7). Equation (7) is applied M times to create the final tree model.

$$F_0(x) = \sum_{i=1}^{n} L(y_i, r) \tag{4}$$

$$r_{im} - \left[\frac{L\left(y_i, F(x_i)\right)}{F(x_i)} \right] F_{(x)} = F_{m-1}(x) \tag{5}$$

$$r_m = \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + r^h m(x_i)\right) \tag{6}$$

$$F_0(x) = F_{m-1}(x) + r_m h_m(x) \tag{7}$$

The machine learning techniques of SVM, RF, and GBRT used in this study were reviewed. Since SVM, RF, and GBRT can be applied not only to classification but also to regression, they can be used to predict fire damage. The advantages and disadvantages of the three techniques are next described. SVM has the advantage of not being greatly affected by noise, overfitting does not occur easily, and its accuracy is high. However, to find the optimal model, a combination of hyperparameters must be tested, and if there are many observations and attributes in the input data, it takes a long time to learn, the result is difficult to interpret, and it is not known how the result was obtained. RF can avoid overfitting by the law of large numbers, is fast, has excellent accuracy, and is not greatly affected by noise or outliers. However, its hyperparameters must be selected carefully and it is relatively slow with a high memory requirement. GBRT has excellent performance among the supervised learning algorithms, it is unnecessary to adjust the scale of features, and it works well for continuous features. However, the hyperparameters of the model must be carefully selected and the model does not work well for high-dimensional data. Our analysis was conducted by applying these three techniques, which are relatively accurate compared with other machine learning methods and are appropriate for the subject of this study.

3. Experimental Results

3.1 Fire outbreak data collection and processing

The number of fires that occurred in Seoul over the 10 years from 2010 to 2019 was 59060, representing 13.6% of the number of fires nationwide. From 2010 to 2019, the annual average fire rate in Seoul was 0.09% and showed an increasing trend. We thus analyzed the prediction analysis of the amount of property damage occurring during fire accidents in Seoul, which can have a major impact on citizens. In addition, we analyzed the relative importance of several factors related to fire occurrence. The fire data used were disclosed by the National Emergency Management Agency.[2] All the fire data from 2017 to 2019 were disclosed as public big data. However, controversy has arisen over the invasion of privacy and the extent of public data disclosure, and for this reason, only summary data have recently been issued regarding the status of fires. Therefore, the fire data used in this study comprise a total of 18227 fires that occurred in Seoul between 2017 and 2019 (5978 in 2017, 6368 in 2018, and 5881 in 2019). The numbers of casualties were 283 in 2017, 360 in 2018, and 398 in 2019, giving an increase of 40% in 2019 compared with 2017. The amount of property damage increased by 496% from 2017 to 2019. These statistics are summarized in Fig. 2 and indicate that property damage has increased rapidly relative to the numbers of fires and human casualties, thus justifying the need for our study.

The fire data used in this study were from fire-related accidents collected by the Korea Fire and Disaster Management Agency. In 2017, 2018, and 2019, 56, 22, and 22 fire-accident-related items were disclosed, respectively. Investigation items include the fire serial number, dispatch fire station and 911 safety center, date and time, weather conditions, fire location, ignition condition, number of building floors, ignition source, facility classification, structural material, and reception and dispatch times. They also include the distance from the origin of the fire, the

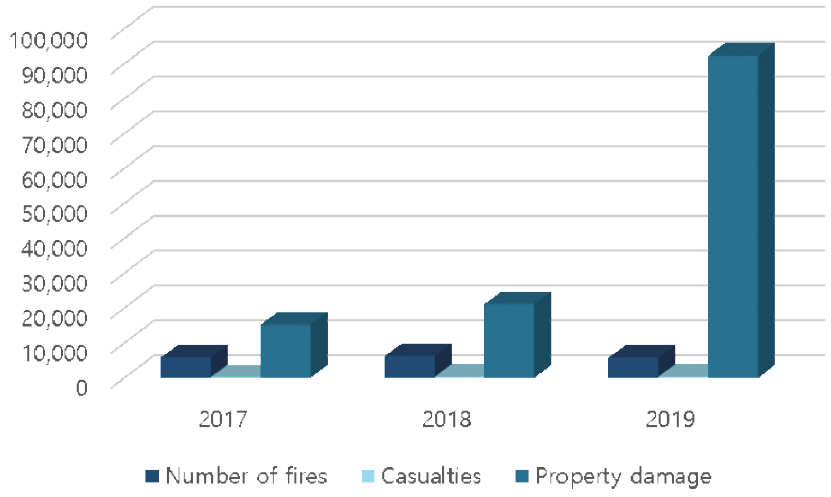


Fig. 2. (Color online) Recent statistics for fires in Seoul, Korea.

number of casualties, property damage, lifesaving status, mobilized manpower and equipment, insurance subscription, and so forth. Among the investigation items provided by the fire accident data, 15 items related to fire occurrence were selected as analysis variables: date and time (month, day, hour, and minutes), fire property damage, weather conditions (wind speed, temperature, humidity, and wind direction), fire location, number of burning floors, facility classification, ignition source, ignition condition, and structural material, as shown in Table 2.

The data were provided as an Excel file, which included many missing values. Since not all the missing values can be eliminated because the reliability would decrease when they were analyzed by applying machine learning techniques, in this study, missing values were treated by a multiple imputation method without removing them. The multiple imputation method predicts the missing values using the remaining variables. This method provides reliable working values through several iterations. Subsequently, we processed categorical data. Among the 15 selected fire variables, date, wind direction, fire location, facility classification, ignition source, ignition condition, and structural material are applicable. Categorical variables were created by converting them into numerical data. The unstructured data were converted into numerical data through the R program. The day of the week of the fire was changed to a number from 1 (Monday) to 7 (Sunday). Wind directions of north, northeast, east, southeast, south, southwest, west, and northwest were changed to values from 1 to 8. Wind speeds of 0–4 m/s and so forth were changed to values from 1 to 3. The fire locations of buildings, structures, and so forth were changed to numbers from 1 to 5. Facility classification included apartment houses, detached houses, health facilities, and so forth, and given values from 1 to 15. Ignition sources, such as flame, spark, and an operating device, were given numbers from 1 to 9. Ignition conditions, such as careless, mechanical, and chemical factors, were given values from 1 to 11. The structural materials, such as reinforced concrete, wood, brick, and block, were given numerical values from 1 to 12.

Table 2
Fire occurrence variables.

Variable	Description
Month	Month of fire
Day	Day of fire
Hour	Hour of fire
Minutes	Minutes of fire
Property damage	Property damage of fire
Wind speed	Wind speed on day of fire
Temperature	Temperature on day of fire
Humidity	Humidity on day of fire
Wind direction	Wind direction on day of fire
Fire location	Buildings and structures, automobiles, railroads, garbage area, etc.
Number of burning floors	Number of burning floors in building
Facility classification	Apartment houses, detached houses, health facilities, public institutions, schools, sanitation facilities, power generation facilities, etc.
Ignition source	Flame, spark, unknown, operating device, firecracker, chemical fire, etc.
Ignition condition	Careless factors, electrical factors, mechanical factors, chemical factors, etc.
Structural material	Wood, reinforced concrete, brick, block, stone, steel frame, etc.

3.2 Adoption of fire accident occurrence factors

We conducted correlation and iterative regression analyses to predict the occurrence of fires in Seoul. Through this, significant variables were found, and the analyses were conducted excluding variables that caused problems of multicollinearity. First, the correlation between fire accident occurrence and each variable was identified. As a result, highly correlated variables were analyzed, and we found that multicollinearity occurs. To resolve this issue, we performed stepwise regression analysis. From the 15 fire variables, seven of them (minutes, day, wind speed, wind direction, humidity, ignition source, and number of building floors) with high multicollinearity were excluded, and the other eight variables were selected (month, property damage, hour, temperature, fire location, facility classification, ignition condition, and structural material) to perform the multiple regression analysis. The suitability of the variables used in the multiple regression analysis was confirmed. The adequacy of the variables was verified through the Durbin–Watson method. The resulting value was 1.860, which is close to 2, indicating that the variables used were suitable for the regression model. By regression analysis, we obtained R^2 of 0.752. Therefore, our model was concluded to have an explanatory power of 75.2%. All eight variables selected were rejected at the significance level of 0.05. We proposed the null hypothesis that the regression coefficient of the independent variable has no direct relationship with the behavior of the dependent and independent variables. Therefore, the independent variables included in these models well explain the variance of the dependent variable. We thus verified the suitability of the multiple regression model and the significance of the regression coefficient, and both were found to be significant.

3.3 Evaluation of predictive power of machine learning techniques for fire accident factors

To predict and analyze the risk factors of fire accidents in Seoul, the fire occurrence data were divided into training and test data in the ratios of 7:3 and 8:2, and the model with the highest significance was used in the R Studio program.[25] Furthermore, since the cost of the prediction accuracy must be compared between the models and analyzed under the same conditions, the same training and test data were used for all models. Machine learning methods such as SVM, RF, and GBRT were compared and analyzed. Accuracy was also compared and analyzed through the MAE and RMSE. After evaluating the entire data set of Seoul using the three methods, a model with high predictive power was adopted. In addition, we predicted and analyzed the amount of fire property damage through a model with high predictive power, and we also compared and analyzed the actual data and the predicted amount of fire property damage. Finally, to analyze the factors that affect the occurrence of fires, we evaluated the risk factors for fire accidents, focusing on the variables that cause them.

3.3.1 Comparative analysis of machine learning cross-validation predictive power

Using k-fold cross-validation, the MAE and RMSE values of the validation and test data were derived for each ratio, and the final model with the smallest value was selected, as shown in Tables 3 and 4. A total of 8642 training data and 3703 test data were analyzed for the 7:3 ratio, and 9876 training data and 2469 test data were analyzed for the 8:2 ratio. The comparative analysis of the results of validation and test data values for ratios of 7:3 and 8:2 confirmed that the validation and test data of 8:2 have smaller MAE and RMSE values, indicating their higher predictive power. Furthermore, the MAE and RMSE values of the test data for the 8:2 ratio are smaller than those of the validation data, which means that the predictive power of the test data is higher. Therefore, as a result of the comparative analysis of MAE and RMSE for SVM, RF, and GBRT, the RF model was found to have the highest predictive power, followed by GBRT then SVM. The RF model with the highest predictive power had the smallest MAE and RMSE at 450. The validation data had an MAE value of 2.027 and an RMSE value of 2.428, and the test data had an MAE value of 2.013 and an RMSE value of 2.057. When the number of trees was 450, the differences between the values of validation and test data were 0.01 for the MAE value and 0.3 for the RMSE value; thus, the level of overfitting was found to be relatively low.

3.3.2 Prediction and evaluation of amount of fire damage

As a result of the machine learning analysis of the data for three years (2017–2019) of fire accidents in Seoul, RF was found to be the most significant model when the ratio of validation data to test data was 8:2. Therefore, RF, which was judged to have the highest predictive power

Table 3
Comparative analysis of predictive power of each model for Seoul data with ratio of 7:3.

		Validation data		Test data	
		MAE	RMSE	MAE	RMSE
SVM	$C = 4, \gamma = 0.2, \varepsilon = 0.01$	5.572	6.739	5.468	6.987
	$C = 4, \gamma = 0.2, \varepsilon = 0.02$	5.441	6.692	5.408	6.913
RF	Estimators = 250	2.452	2.920	2.378	2.619
	Estimators = 450	2.238	2.693	2.148	2.194
GBRT	Estimators = 400	3.034	3.798	3.108	3.797
	Estimators = 1000	3.012	3.715	3.035	3.648

Table 4
Comparative analysis of predictive power of each model for Seoul data with ratio of 8:2.

		Validation data		Test data	
		MAE	RMSE	MAE	RMSE
SVM	$C = 4, \gamma = 0.2, \varepsilon = 0.01$	5.313	6.554	5.101	6.739
	$C = 4, \gamma = 0.2, \varepsilon = 0.02$	5.134	6.303	4.948	5.613
RF	Estimators = 250	2.190	2.604	2.081	2.348
	Estimators = 450	2.027	2.428	2.013	2.057
GBRT	Estimators = 400	2.943	3.498	3.067	3.683
	Estimators = 1000	2.821	3.283	2.985	3.448

among the machine learning techniques in this study, was used to predict the amount of fire property damage, and, using this model, the relative importance of variables was analyzed. On the basis of the previous results, the analysis focused on the test data with the ratio of validation data to test data of 8:2 assessed by RF. However, in the case of property damage, there are problems with outliers because the range is large. We therefore conducted the analysis after removing the outliers. The outliers were identified through the generally accepted interquartile range (IQR) rule. From a total of 3645 data, 2430 cases were analyzed. As a result, the performance of the predictive model was evaluated using the R-squared technique, and the predicted data showed an accuracy of 83%. The statistical analysis results for the amount of damage are shown in Table 5.

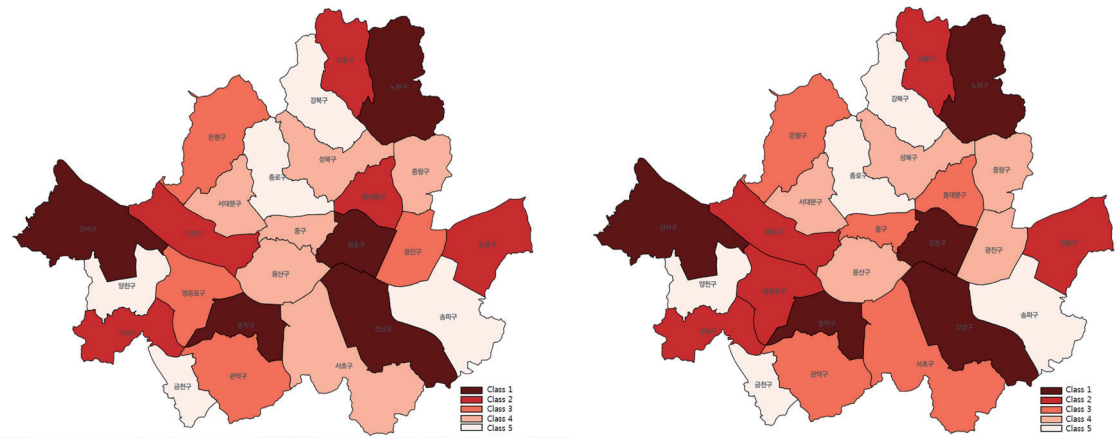
A spatial analysis based on five classes of the amount of fire damage was performed on the predicted amount of fire damage, the result of which is shown in Fig. 3. Here, Class 1 is the area with the most fire property damage and Class 5 is the area with the least fire property damage.

After confirming the results through the spatial analysis of actual and predicted data, we conducted importance analysis to evaluate the factors of fire damage centering on the top 10% of fires with the highest amount of property damage. When we analyzed the importance of fire occurrence factor variables using RF, focusing on 240 cases with high fire property damage, we found that the facility classification had the highest correlation, followed by fire location, ignition condition, temperature, hour, month, and structural material, as shown in Table 6.

Using these seven variables and their importance levels, we conducted a more detailed analysis. The most important variable was facility classification, which includes apartment

Table 5
Prediction of fire property damage.

	Average	Median	Minimum	Maximum
Actual data	2146	134	0	483000
Predicted data	2332	120	18	422344



(a) (b)

Fig. 3. (Color online) Fire property damage prediction map. (a) Actual data. (b) Predicted data.

Table 6
Evaluation of importance of fire factors by RF.

Variable	Importance level (%) for tree number 450
Facility classification	29.78
Fire location	22.16
Ignition condition	18.41
Temperature	11.19
Hour	8.67
Month	6.23
Structural material	3.56

houses, detached houses, health facilities, public institutions, schools, sanitation facilities, and power generation facilities. Among these, fires occurred most commonly in apartment houses, followed by restaurants, detached houses, and service facilities. Fire locations included buildings and structures, automobiles, railroads, and garbage areas. For this variable, fire was most prevalent in buildings and structures. The ignition condition includes elements such as construction, dangerous goods, automobiles, and garbage fires. For this variable, most fire accidents were due to carelessness, followed by electrical and mechanical factors. For the variable of temperature, we divided the temperature into 5 °C units for analysis and found that the most fire accidents occurred in the temperature range of 25 to 29.9 °C, followed by the range of 5 to 9.9 °C. For the variable of hour, the most frequent occurrence of fires was between 6:00 pm and 9:00 pm. For the variable of month, the year was divided into four quarters for analysis, and it was found that the most fire accidents occurred in the third quarter (July to September). Finally, among the structural materials, fire accidents were found to occur most often in brick structures.

4. Conclusions

In this study, we used three years of fire big data for Seoul. The factors contributing to future fires were predicted by applying machine learning techniques, using SVM, RF, and GBRT for the analysis. In addition, k-fold cross-validation was used to increase the predictive power of machine learning. The accuracy of the technique was evaluated through MAE and RMSE. In addition, the amount of fire property damage was predicted using the model judged to have the highest predictive power among the machine learning techniques, and using this model, the importance of fire variables affecting fire occurrence was analyzed and evaluated.

First, variables were acquired for fire prediction analysis. Among the variables provided by the National Fire Administration of Korea, the month, hour, temperature, fire location, facility classification, ignition condition, and structural material were selected as variables with no multicollinearity through correlation and multiple regression analyses, and machine learning was performed using them.

Second, for predictive analysis, MAE and RMSE were calculated using validation and test data for different ratios of validation data to test data. The analysis also used k-fold cross-validation based on SVM, RF, and GBRT to evaluate validation and test data. The best results

were obtained with RF and a ratio of validation data to test data of 8:2, followed by GBRT and SVM. Furthermore, in the case of RF, when the number of trees was 450, the difference between the MAE and RMSE values of the validation and test data was the smallest; thus, the problem of overfitting was also the lowest.

Third, we performed a comparative analysis of the amount of fire property damage using RF, and the accuracy of the predicted data was 83%. An analysis was conducted on the factors of fire damage, focusing on data for a high amount of fire property damage. For the facility classification variable, which includes apartment houses, detached houses, health facilities, public institutions, schools, sanitation facilities, and power generation facilities, we found that most fires occurred inside apartment houses followed by restaurants, detached houses, and service facilities.

In summary, the amount of fire property damage was estimated using fire data for Seoul, and machine learning techniques were applied to the data for high amounts of damage. This study is expected to provide guidelines for analyzing important factors that affect the occurrence of fire accidents. This will enable the spatial distribution of fire accidents to be determined and the establishment of corrective measures to manage and reduce urban fires as well as reduce the loss of life and damage to property.

Acknowledgments

This research was supported by a grant (2021R1F1A106422811) from the Basic Research Project for Science and Engineering, funded by the Ministry of Science and ICT of the Korean government.

References

1 E. P. Lee: A Study on Eire Data Analysis in Korea, Japan and USA (2) Direct Property Losses Due to Fires **18** (2004) 4. https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO200416642219884&SITE=CLICK
2 National Fire Data System: https://www.nfds.go.kr/index.do (accessed July 2021).
3 National Fire Protection Association: https://www.nfpa.org/ (accessed September 2021).
4 H. H. Yun, K. Y. Baek, and B. H. Park: J. Korean Soc. Hazard Mitig. **1** (2001) 123. http://www.koreascience.or.kr/article/JAKO200117821847167.page
5 Y. S. Kang: Plan. Asso. **38** (2003) 65. https://kpa1959.or.kr/?menuno=
6 P. Chhetri, J. Corcoran, R. J. Stimson, and R. Inbakaran: Geo. Res. **48** (2010) 75. https://doi.org/10.1111/j.1745-5871.2009.00587.x
7 C. Hastie and R. Searle: Fire Safety **84** (2016) 50. https://doi.org/10.1016/j.firesaf.2016.07.002
8 C. R. Jennings: Fire Safety **62** (2013) 13. https://doi.org/10.1016/j.firesaf.2013.07.002
9 I. A. Lee, H. R. Oh, and Z. K. Lee: Bigdata **6** (2021) 133. https://doi.org/10.36498/kbigdt.2021.6.1.133
10 M. Madaio, S. T. Chen, O. L. Halimson, W. Zhang, and X. Cheng: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (ACM SIGKDD, 2016) 185–194.
11 N. M. Al-Abdaly, S. R. Al-Taai, H. Imran, and M. Ibrahim: Enterp. Technol. **5** (2021) 59. https://doi.org/10.15587/1729-4061.2021.242986
12 E. J. M. Carranza and A. G. Laborte: Comput. Geosci. **74** (2015) 60. https://doi.org/10.1016/j.cageo.2014.10.004
13 M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan: Comput. Geosci. **148** (2021) 1. https://doi.org/10.1016/j.cageo.2021.104688
14 T. W. Chu, A. Shirmohammadi, H. Montas, and A. Sadeghi: Am. Soc. Agric. Eng. **47** (2004) 1523. https://doi.org/10.13031/2013.17632

15 W. C. Wang, K. W. Chau, C. T. Cheng, and L. Qiu: Hydrol. J. **374** (2009) 294. https://doi.org/10.1016/j.jhydrol.2009.06.019
16 S. S. Band, S. Janizadeh, S. C. Pal, I. Chowdhuri, Z. Siabi, A. Norouzi, A. M. Melesse, M Shokri, and A. Mosavi: J. Sens. **20** (2020) 5763. https://doi.org/10.3390/s20205763
17 V. Vapnik and A. Lerner: Autom. Remote Control **24** (1963) 774. https://cir.nii.ac.jp/crid/1571135650527018624
18 C. S. Yu, C. J. Lin, and J. K. Hwang: Protein Sci. **13** (2004) 1402. https://doi.org/10.1110/ps.03479604
19 L. Xing, W Yan, and J. Zhicheng: J. Syst. Simul. **33** (2021) 2606. https://doi.org/10.16182/j.issn1004731x.joss.21-FZ0705
20 L. Breiman: Mach. Learn. **45** (2001) 5. https://doi.org/10.1023/a:1010933404324
21 Z. Fei, F. Yang, K. L. Tsui, L. Li, and Z. Zhang: J. Energy **225** (2021) 1. https://doi.org/10.1016/j.energy.2021.120205
22 L. Breiman: Classification and Regression Trees, Chapman & Hall/CRC (Routledge, New York, 1984) 1st ed., pp. 59–218
23 Z. Liu, G. Gilbert, J. M. Cepeda, A. O. Lysdahl, L. Piciullo, H. Hefre, and S. Lacasse: Geosci. Front. **12** (2021). https://doi.org/10.1016/j.gsf.2020.04.014
24 J. H. Friedman: Ann. Statist. **29** (2001) 5. https://doi.org/10.1214/aos/1013203451
25 S. Y. Hu, J. Y. Kim, and T. H. Moon: J. Korean Assoc. Geogr. Inf. Stud. **4** (2018) 21. https://doi.org/10.11108/kagis.2018.21.4.064

About the Authors

Min Song Seo received her B.S. and M.S. degrees from Gyeongsang National University, Republic of Korea, in 2016 and 2018, respectively. She is currently working on her Ph.D. degree. Her research interests are in GIS analysis using big data. (msong7938@gmail.com)

Ever Enrique Castillo-Osorio received his B.S. degree from Inca Garcilaso de la Vega University, Peru, in 1999 and his M.S. and Ph.D. degrees from Gyeongsang National University, Republic of Korea, in 2017 and 2022, respectively. From 2000 to 2015, he worked on ICT and GIS at the Meteorology and Hydrology Service of Peru. He is currently a researcher at Yonsei University. His research interests include GIS, machine learning, and disaster risk management. (ever.castillo.osorio@gmail.com)

Hwan Hee Yoo received his B.S. degree from Kangwon National University, Republic of Korea, in 1981 and his M.S. and Ph.D. degrees from Yonsei University, Republic of Korea, in 1983 and 1988, respectively. Since 1990, he has been a professor at Gyeongsang National University, Korea. He served as the president of Korean Society for Geospatial Information Science from 2009 to 2010. His research interests are in GIS and big data analysis.
(hhyoo@gnu.ac.kr)