Sensors and Materials, Vol. 34, No. 12 (2022) 4841–4854
MYU Tokyo

S & M 3147

Analysis of Fire Risk Factors in Seoul, Korea, by Machine Learning

Min Song Seo, Ever Enrique Castillo Osorio, and Hwan Hee Yoo*

Department of Urban Engineering, Gyeongsang National University,
501, Jinju daero, Jinju, Gyeongsangnam-do 660701, Korea

(Received April 27, 2022; accepted August 9, 2022; online published September 7, 2022)

Keywords: fire accidents, support vector machine, random forest, gradient boosted regression tree, mean absolute error, root mean squared error

Different types of fire accidents in the urban area of Seoul, Korea are continuously occurring, causing risk and damage to property and life. In this study, we analyze various spatial and non-spatial fire risk factors by applying machine learning techniques to predict their level of importance in future events. We use the data on fire accident for three years (2017–2019) published by the Korean Fire Service and the Seoul Metropolitan Government. Regarding the machine learning techniques, we use support vector machine (SVM), random forest (RF), and gradient boosted regression tree (GBRT). As the first phase, a multiple regression analysis is performed to select seven main factors related to fire occurrence. In the second phase, we calculate the mean absolute error (MAE) and root mean squared error (RMSE) using validation and test data for the machine learning techniques, revealing that RF obtains ideal results. In the third phase, we analyze the importance of the seven fire factors using RF, resulting in the ignition condition (produced by electrical, mechanical, and chemical reasons) being the main factor in fire occurrence. This study is expected to be used as an important guideline to define urban fire reduction and management measures in Seoul, the capital of South Korea.

1. Introduction

In Korea, fire accidents are one of the frequent disasters along with traffic accidents. Fire accidents can happen at any time and place. Therefore, it is important to prevent them and take countermeasures, since their risk is frequently increasing. In 2020, there were 38659 fire accidents in Korea, resulting in 2282 casualties and an estimated $479.44 million in property damage.[1] Specifically, the number of fire accidents in Seoul, the capital of South Korea, has risen and fallen cyclically, with 5978 cases in 2017, 6368 cases in 2018, and 5881 cases in 2019. However, in the case of casualties, there were 283 people involved in 2017, 360 in 2018, and 398 in 2019, increasing 40% in 2019 compared with 2017. In addition, the amount of property damage increased by 496% in 2019 compared with 2017, estimated at $12.21 million in 2017, $16.92 million in 2018, and $73.77 million in 2019 according to reports from the National Fire

*Corresponding author (ERI): e-mail: hhyoo@gnu.ac.kr
https://doi.org/10.18494/SAM3955

ISSN 0914-4935
https://myukk.org/

Data System (NFDS).[1] Moreover, since a large amount of damage occurs in the urban areas of Seoul, urgent measures are needed to prevent and reduce fire accidents. In Korea, fire accidents produce considerable damage to property and life; therefore, the NFDS has been in operation since 2007 to prepare countermeasures. The NFDS is a database that contains fire event information. Since 2009, the website allows the public to review statistical data on fire accidents in terms of their current status, classification by region, administrative division, and damage report. However, in the case of NFDS, since only basic statistics are provided, it is insufficient to understand the in-depth risk level.

Large and small fires causing damage in urban areas happen continuously not only in Korea but also abroad as reported by Lau *et al.*[2] and the National Fire Protection Association of the United States.[3] As the risk of fire accidents increases in urban areas around the world, various studies are actively being conducted. Zhang defined the term 'urban fire risk' by referring to both the definition of risk and the characteristics of urban fires in the International Strategy for Disaster Reduction enacted by the United Nations (UN) in 2004.[4] Additionally, the urban fire risk was comprehensively analyzed in Haikou, the capital of Hainan Province, China. The gray correlation degree method was applied on the basis of data from statistics of past fire accidents. In addition, the weight factor of the system was set and the analytic hierarchy process (AHP) was used on the basis of the quantitative analysis of statistical indicators, and through this, an urban fire risk assessment system was established.[4] Xin and Huang[5] presented a building fire risk analysis model based on scenario clusters and a method applied to fire risk management in buildings. In the building fire risk analysis, the average fire risk of a residential building was quantified by constructing a scenario cluster and selecting the number of fatalities and property losses as fire risk indicators. On the basis of analysis results, the fire safety rating in buildings was improved by using the appropriate fire risk model type. Hastie and Searle[6] confirmed through the analysis of data provided by the West Midlands Fire Service (WMFS) of the UK that serious fires occurred continuously from September 2010 to August 2013. The authors performed a regression analysis on the occurrence of fires in residential areas, considering the number of fires as the dependent variable and the socioeconomic and demographic data as independent variables. As a result, the largest number of fires was identified as occurring in places where a large number of black people lived, single-person households under the age of 65, and those who had not worked for more than five years. Song *et al.*[7] plotted the frequency-damage and rank-damage distributions of fire accidents in cities of China and Switzerland, and verified whether the power-law relationship was appropriate. This was confirmed by plotting the distributions of frequency magnitude (loss) and rank magnitude (Zipf plot) of urban fires. As a result, the fire data of the two countries had a useful power-law distribution. This power-law relationship did not change with scale and time. In other words, it was confirmed that both countries fit the exponential relationship well and are constant regardless of size and time. Telesca and Song[8] analyzed the temporal distribution of urban fires from 2000 to 2009, focusing on the fires in Anshan, China. As a result, it was revealed that the core process of city fire is a fractal process with a high degree of time clustering of events. In addition, time tends to cluster as losses increase after a fire. The application of multiple fractal trend variability analysis to urban fires suggested that the sequences are dynamically

heterogeneous owing to different long-range temporal correlation properties for variability between large and small events. Lu *et al.*[9] studied the frequency-damage distribution and time-scaling characteristics for fires that caused more than three fatalities between 2002 and 2009. Six factors, namely, place, cause, time, season, year, and fire area, were analyzed to evaluate their impact on the fire, and a scaling index was used for comparative analysis. Factors such as nonresidential locations, electrical causes, winter months, and regions with strong economies have slowed down the fire frequency as the number of deaths increased. That is, the time scale index was found to decrease significantly as the number of deaths increased. Rohde *et al.*[10] conducted a predictive analysis of residential home fires across southeast Queensland (SEQ) in Australia by applying Bayesian methodology. The number of fires expected in a year was calculated for SEQ areas where the expected risk ranged from less than two cases per year to a maximum of 25 cases per year. In the service level agreement (SLA), it was analyzed that there was a high correlation between the population size and the number of fires, as well as between the number of buildings and the number of fires. These results were presented as a distribution map on the geographic information system (GIS), concluding that many of the fires occurred in residential areas. In addition, various fire-related studies such as fire risk analysis, building fire risk analysis, and fire correlation analysis are being conducted. However, there is a lack of research on predicting factors that cause fires and predicting future fires through them.

The purpose of this study is to predict the risk of spatial and non-spatial factors (e.g., location, time, cause, and climate) that most affect the different types of urban fires through the use of machine learning techniques. A predictive analysis of fire occurrence factors was conducted for Seoul Metropolitan City using data provided by the Korea Fire Service for three years (2017–2019). For the analysis, multiple regression analysis (MRA) models and machine learning techniques such as support vector machine (SVM), random forest (RF), and gradient boosted regression tree (GBRT) were used. The accuracy of each model was presented through the mean absolute error (MAE) and root mean squared error (RMSE), obtaining a model with a high degree of adjustment. The flowchart of the processes considered in this study is illustrated in Fig. 1.

2. Methodology

2.1 SVM

The SVM algorithm is a nonlinear generalization algorithm and has become a solid foundation for statistical learning theory.[11] The basic SVM is widely used in binary classification problems. One side is divided into a positive class and the other side into a negative class centering on the hyperplane.[12] The basic idea of SVM is to find a hyperplane that can properly separate the data constituting two categories (positive and negative classes). SVM aims to find the hyperplane with the maximum distance to the nearest sample point when used for classes.[13]

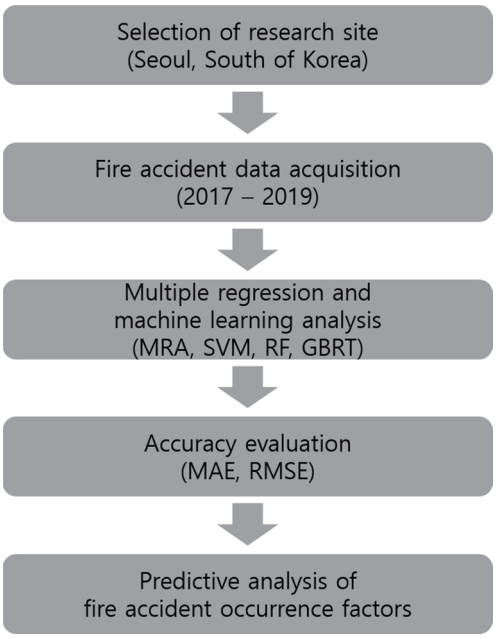

Fig 1. Flowchart of the study.

2.2 RF

The RF algorithm uses the bootstrap method as an ensemble learning model. This is a method used to generate multiple samples and apply a decision tree model to synthesize the obtained results.[14] RF combines multiple decision tree models and builds the decision tree using bootstrap training data in the same way as bagging. The RF training process constructs several different decision trees. A random subset is taken from each split of the tree node for analysis. Furthermore, each tree is generated from a randomly selected subset of training samples via bootstrap sampling. Bootstrap sampling randomly selects m samples by substituting from a training set with n observations. Because this randomness is introduced, RF can be used to analyze more patterns in the data by increasing tree diversity.[15]

2.3 GBRT

The GBRT algorithm is an ensemble method based on the classification and regression tree (CART) algorithm.[16] GBRT is a model that combines two techniques: boosting and regression. This combination improves the model accuracy and reduces the variance. Similarly to other boosting methods, GBRT trains multiple CART basic learners through multiple iterations and ultimately produces strong learners with a linear combination of these weak learners. GBRT as well as RF can be applied to both classification and regression, with high accuracy and no need for scaling. In addition, it has the advantage that it works well for continuous characteristics.[17]

2.4 Predictive evaluation using machine learning techniques

In algorithm discovery and modeling in the predictive analysis of fire occurrence using machine learning, an algorithm for practical operation is selected and a decision is made on how to model the algorithm. The criterion for selecting algorithms and modeling methods is their applicability. Therefore, it is necessary to review how to judge this applicability. For the use of machine learning in predicting fire risk, the accuracy of the machine learning model must be recognized. The MAE and RMSE of each model are frequently compared to evaluate the accuracy of a numerical prediction model. In addition, they can be used to evaluate the accuracy of machine learning models. RMSE is one of the widely applied error index statistics.[18] It is generally accepted that the lower the RMSE value, the higher the model efficiency. This limits what is considered a low RMSE and is based on the standard deviation of observations.[19] Moreover, MAE is another error indicator frequently used in model evaluation. A value of 0 indicates an ideal adjustment. The lower the RMSE and MAE values of the calculated data, the better the model evaluation.[20] MAE is determined using Eq. (1), where it is the average of the absolute error between the measured and actual values. In addition, Eq. (2) shows the formula for RMSE, which represents a value obtained by calculating the root mean square of the error between the measured and actual values.[15]

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}\left(\widehat{y_i} - y_i\right)^2} \tag{2}$$

To increase the accuracy of the machine learning techniques, it is necessary to derive a global optimum that can explain the data distribution. 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.[21,22] 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 of machine learning.[21] The optimal hyperparameter can be derived through various trial and error tests. Therefore, in our work, a *k*-fold cross validation method was used.[23] The *k*-fold cross validation divides the training data into *k* equal parts, uses *k*–1 of the *k*-divided training data for setting the search range of the hyperparameters, and validates the model performance using the remaining training data as validation data. The validation process is repeated *k* times to determine the hyperparameter with the lowest generalization error as the final model to which the test data is applied. Through this, the error correction was greatly improved by selecting the ideal model and collecting the thresholds and weights as initial values. The verification process was repeated *k* times, and the hyperparameter with the lowest generalization error was appointed as the final model. The training and test data were divided in a 7:3 ratio and cross-verified

through the *k*-fold cross validation method. The test data was applied to the hyperparameter with the lowest MAE and RMSE of the validation data for each model. These processes were analyzed through the R Studio program. In addition, since the prediction accuracy between the models was comparatively analyzed, the same training data was used for all models.

3. Experimental Results

3.1 Data analysis

The number of different types of fires that happened in Seoul, the capital of Korea, over the past 10 years (2010–2019) was 59060, accounting for 13.6% of the total number of fires in the country. Likewise, 2663 injuries (including 363 deaths) were reported in Seoul, accounting for 12.3% of the casualties nationwide. Additionally, the percentage of property damage due to this hazard in Seoul was 4.4% of the total property damage that occurred nationwide. From 2010 to 2019, the average annual increase in fire rate in Seoul was 0.09%, revealing an increasing trend, as presented in Fig. 2.

In our study, we identified several fire factors that cause fire accidents in Seoul to analyze which of them act as important factors in the occurrence of this hazard. We used the fire data published by the Korea National Fire and Disaster Prevention Administration. From 2017 to 2019, all fire data were released as public big data. However, owing to controversy over the invasion of privacy and the scope of public data disclosure, the latest data displays a summary of the fire status. Therefore, we used in this study the fire data of Seoul from 2017 to 2019, with a total of 18227 cases. The number of fires rose and fell cyclically, with 5978 cases in 2017, 6368 cases in 2018, and 5881 cases in 2019. However, the numbers of victims were 283 in 2017, 360 in 2018, and 398 in 2019, reaching a 40% increase in 2019 compared with 2017. Additionally, the amount of property damage increased by 496% from 2019 compared with 2017.

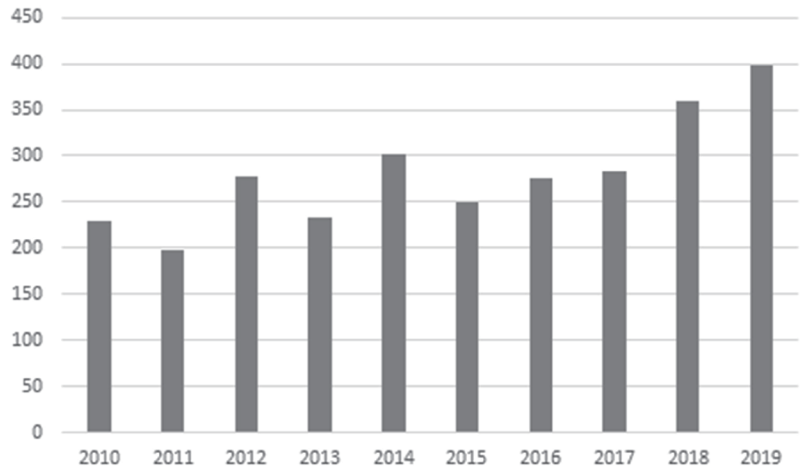

Fig. 2. Casualties in Seoul fire accidents.

Regarding the items related to fire accidents, 56 were disclosed in 2017, 22 in 2018, and 22 in 2019. Investigation items include fire serial number, dispatch fire department and 911 safety center, date and time, weather condition, fire location, ignition condition, number of building floors, ignition source, facility classification, structural material, reception and dispatch time, distance of the fire department and 911 safety center, casualties, property damage, lifesaving status, mobilized personnel and equipment, and insurance purchases. Therefore, 14 elements related to fire occurrence were selected and set as analysis factors. These elements are date and time (month, day, hour, and minutes), weather conditions (wind speed, temperature, humidity, and wind direction), fire location, number of building floors, facility classification, ignition source, ignition condition, and structural material as shown in Table 1.

The analysis of the data for each main fire factor showed that the period with the highest number of fire accidents was in January, and fire accidents occurred more frequently between 1:00 p.m. and 3:00 p.m. and between 6:00 p.m. and 8:00 p.m. In addition, in the case of days of the week, it is found that the rate of fire accidents is highest on Friday and Saturday. In terms of the facility classification, the highest rate of fire accidents happened in residential facilities, followed by living service and business facilities. In the case of the ignition source, the highest number of fire accidents occurred owing to carelessness, followed by electrical and mechanical reasons.

3.2 Correlation analysis of fire accident factors

Correlation and iterative regression analyses were applied to find the most significant factors to carry out fire occurrence prediction analysis in Seoul. Likewise, factors that cause multicollinearity problems were excluded. First, the correlation between the occurrence of fire accidents and each factor was examined. Factors with high correlation and possibility of multicollinearity events were obtained. To solve this problem, a stepwise regression analysis was

Table 1
Fire occurrence factors.

Factor	Description
Month	Month of fire
Day	Day of fire
Hour	Hour of fire
Minutes	Minutes of fire
Wind speed	Wind speed on the day of fire
Temperature	Temperature on the day of fire
Humidity	Humidity on the day of fire
Wind direction	Wind direction on the day of fire
Fire location	Buildings and structures, automobiles and railroads, garbage area, etc.
Number of building floors	Number of burning floors in the 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	Carelessness, electrical reasons, mechanical reasons, chemical reasons, etc.
Structural material	Wood, reinforced concrete, brick, block, stone, steel frame, etc.

performed. As a result, seven fire occurrence factors with high multicollinearity (minutes, day, wind speed, wind direction, humidity, ignition source, and number of building floors) were excluded, and multiple regression analysis was performed with the remaining seven factors (month, hour, temperature, fire location, facility classification, ignition condition, and structural material). Subsequently, the suitability of the factors used in the multiple regression analysis was determined by the Durbin–Watson method. Since the result is 1.860, which is close to 2, the factors are considered suitable for the regression model. As a result of the regression analysis, R^2, which can confirm the explanatory power of the model, was 0.835, that is, 83.5%. In addition, all the standardization factors with the significance level $\alpha > 0.05$ were rejected. Regarding the regression coefficient of the factors, it is assumed in the null hypothesis that the dependent and independent variables are different from each other. Therefore, the independent variables included in the models can affect the variance of the dependent variable. Table 2 shows the results of the multiple regression model analysis, where B is the Beta coefficient and t is the t-value results of the coefficient.

3.3 Evaluation of machine learning techniques for the prediction of fire accident factors

Machine learning methods such as SVM, RF, and GBRT were comparatively analyzed to conduct the predictive analysis of fire accident risk factors in Seoul. In addition, the accuracy of the obtained results was estimated using their MAE and RMSE. After analyzing all the data of Seoul using three methods, a model with high predictive power was chosen. Moreover, the risk factors for fire accidents were analyzed, with a focus on those that cause these accidents through a model with high predictive power.

3.3.1 SVM

Hyperparameters are important for determining the optimal SVM model. In this study, the hyperparameters were tuned for k-fold cross validation. The hyperparameters are C, γ, and ε. C is related to the error and its effect on the validation data. Likewise, ε is needed to determine the allowable error rate. Basically, by applying the kernel function, the curvature of the hyperplane

Table 2
Results of multiple regression model analysis.

	Non-standardization factor		t-value	Standardization factor
	B	Standard error		
Month	0.267	0.038	3.420	0.000
Hour	0.342	0.024	6.116	0.000
Temperature	−0.543	0.235	−16.217	0.000
Fire location	0.681	0.134	9.421	0.000
Facility classification	0.812	0.138	4.812	0.002
Ignition condition	0.581	0.347	3.546	0.001
Structural material	−0.438	0.526	−10.485	0.000
R^2				0.835
Durbin–Watson				1.860

varies depending on the increasing γ. In addition, the explanatory power of the validation data increases. However, there may be a problem of overfitting; therefore, a model with small γ should be selected as the final model. Since it is related to the effect of validation data, it was selected to be 0.2 through analysis. In the case of ε, it was chosen to be 0.01 or 0.02. As a result of the hyperparameter, the model with the minimum MAE and RMSE was finally selected as the optimal model for SVM. The results of the adjustment showed that MAE (5.172) is the lowest when $C = 4$, $\gamma = 0.2$, and $\varepsilon = 0.02$, and RMSE (5.239) is the lowest when $C = 4$, $\gamma = 0.2$, and $\varepsilon = 0.02$, which are determined as the final hyperparameters. Table 3 shows the results of the SVM adjustment for the fire accident data of Seoul.

3.3.2 RF

RF is used to perform *k*-fold cross validation while changing the number of trees, obtaining a hyperparameter for each case. The combination of the number of trees with the lowest MAE and RMSE among validation data was chosen as the final model. The number of trees ranges from 50 to 500, and analysis was performed in units of 50 intervals. As a result of the analysis, MAE and RMSE do not change significantly in the levels 250 and 450. When the number of trees exceeds 250, MAE is 2.765 and RMSE is 2.820. Moreover, when the number of trees is 450, MAE is 2.498 and RMSE is 2.593, which are minimum values. Therefore, the models with 250 and 450 trees were selected as the final models. These final models were determined by considering the improvements of MAE and RMSE. Table 4 shows the RF adjustment results for fire accident data in Seoul.

3.3.3 GBRT

GBRT defines the number of trees and the learning rate as the main hyperparameters. The learning rate is a hyperparameter that controls the amount of error correction from the previous tree. The learning rate was set to be 0.1 and the upper limit of the number of trees was defined as 1000, starting from 50. Subsequently, the final model was determined considering whether

Table 3
Results of SVM analysis.

Parameter			Output	Error	
C	γ	ε		MAE	RMSE
1	0.2	0.01	Output1	6.145	6.592
1	0.2	0.02	Output2	6.124	6.529
2	0.2	0.01	Output1	5.965	6.312
2	0.2	0.02	Output2	5.948	6.307
3	0.2	0.01	Output1	5.618	5.994
3	0.2	0.02	Output2	5.604	5.826
4	0.2	0.01	Output1	5.248	5.490
4	0.2	0.02	Output2	5.172	5.239
5	0.2	0.01	Output1	5.424	5.797
5	0.2	0.02	Output2	5.475	5.848

Table 4
Results of RF analysis.

Number of trees	Error	
	MAE	RMSE
50	3.648	3.872
100	3.614	3.772
150	3.287	3.498
200	3.045	3.209
250	2.765	2.820
300	2.721	2.811
350	2.703	2.803
400	2.694	2.781
450	2.498	2.593
500	2.527	2.680

Table 5
Results of GBRT analysis.

Number of trees	Error	
	MAE	RMSE
50	5.421	5.647
100	4.834	5.132
150	4.734	5.021
200	4.694	5.011
250	4.642	4.912
300	4.164	4.621
350	4.153	4.536
400	3.334	3.621
450	3.332	3.607
500	3.321	3.593
600	3.231	3.502
700	3.247	3.589
800	3.135	3.438
900	2.949	3.197
1000	2.915	3.015

MAE and RMSE were improved. As a result of the analysis, it was observed that when the number of trees exceeded 400, the decrease ranges of MAE and RMSE were reduced by one or less, until they became the lowest when the number of trees reached the value of 1000. Consequently, a model with 400 trees with critical low reductions in MAE (3.334) and RMSE (3.621) and a model with 1000 trees displaying the lowest data for MAE (2.915) and RMSE (3.015) were selected as the final models, as shown in Table 5.

3.3.4 Comparative analysis of predictive power by machine learning

By *k*-fold cross validation, MAE and RMSE values of validation and test data were derived, and the final model with the smallest MAE and RMSE values was selected, as shown in Table 6. The analysis results revealed that the MAE and RMSE values of the test data are smaller than those of the validation data, which means that the predictive power of the test data is correct. Therefore, the results of the comparative analysis of MAE and RMSE for the machine learning techniques SVM, RF, and GBRT revealed that the RF model has the highest predictive power, followed by the GBRT and SVM models with the highest accuracy. In the RF model results with the highest predictive power, MAE and RMSE were smaller when the number of trees was 450 than when the number of trees was 250. Therefore, the validation data was analyzed as the MAE of 2.498 and the RMSE of 2.593. In the case of the test data, the MAE was 2.448 and the RMSE was 2.694. Through SVM, the greatest difference between the MAE and RMSE values of the validation and test data was obtained. Thus, the level of overfitting can be analyzed as having a relatively high SVM. On the other hand, the results of the RF analysis revealed that if the number of trees is 450, the differences between the MAE and RMSE values of the validation and test data are 0.05 and 0.101, respectively, which indicates that the level of overfitting is relatively low.

Table 6
Comparative analysis of predictive power by model for Seoul data.

		Validation data			Test data	
		MAE	RMSE		MAE	RMSE
SVM	$C = 4$, $\gamma = 0.2$, $\varepsilon = 0.01$	5.248	5.490	Output	5.468	5.687
	$C = 4$, $\gamma = 0.2$, $\varepsilon = 0.02$	5.172	5.239		5.418	5.513
RF	Estimators = 250	2.765	2.820		2.678	2.719
	Estimators = 450	2.498	2.593		2.448	2.694
GBRT	Estimators = 400	3.334	3.621		3.508	3.897
	Estimators = 1000	2.915	3.015		3.168	3.348

Table 7
Evaluation of importance of fire factors by RF.

Factors	Importance level for tree number of 250 (%)	Importance level for tree number of 450 (%)
Month	9.14	9.68
Hour	15.94	15.19
Temperature	2.08	2.12
Fire location	19.21	19.18
Facility classification	20.12	20.24
Ignition condition	28.37	29.20
Structural material	5.14	4.39

3.3.5 Analysis and evaluation of importance of fire factors by RF

Through the comparative analysis of the results in Table 6, it has been revealed that RF is the most significant model. Therefore, the relative importance between factors was analyzed using RF, which was considered to have the highest predictive power among the machine learning techniques selected in this study. The severity analysis identifies the important factors in fire occurrence . As a result of the analysis of the importance of each of the seven spatial and non-spatial factors using RF, the ignition condition had the highest correlation with fire occurrence, followed by facility classification, fire location, hour, month, structural material, and temperature, as shown in Table 7.

A more detailed analysis was performed using topic data in order of importance for the seven spatial and non-spatial factors. First, the ignition condition, which is the most important factor, includes carelessness and electrical and mechanical reasons. The analysis of fire data among these reasons demonstrates that the most frequent fire accidents occurred owing to carelessness, followed by electrical and mechanical reasons. The facility classification factor includes elements such as apartment houses, detached houses, health facilities, public institutions, schools, sanitation facilities, and power generation facilities. Among them, fires occurred generally in apartment houses, followed by detached houses, restaurants, and service facilities. The fire location factor includes buildings and structures, automobiles and railroads, and garbage area. Buildings and structures were the places where the majority of fires happened, followed by the garbage area, and then automobiles and railroads. The next factors are hour and month. The fires happened frequently between 6:00 p.m. and 9:00 p.m. In the case of months, the largest numbers of fires occurred in September, followed by October and August. The

structural material factor includes wood, reinforced concrete, brick, block, stone, steel frame, and so forth. Among these reasons, fire accidents commonly occurred in brick structures, followed by wooden and block structures. Finally, regarding the temperature factor, it was analyzed that the majority of fire accidents occurred when the temperature was between 25 and 29.9 °C.

4. Conclusions

Since fires are the second major urban disaster in Korea after traffic accidents, spatial and non-spatial fire risk factors have been analyzed by applying machine learning techniques to fire accident reports. The data of three years (2017–2019) were collected from the Korean Fire Service and the Seoul Metropolitan Government. Among the machine learning techniques, SVM, RF, and GBRT were selected, and their accuracy was evaluated by calculating MAE and RMSE. Furthermore, the obtained results were obtained by analyzing the importance of fire factors that affect the occurrence of this hazard in Seoul.

As the first phase, 14 factors related to fire occurrence were selected among the spatial and non-spatial fire factors disclosed by the Korea Fire Service, and the multicollinearity was analyzed. By stepwise regression analysis, seven factors (minutes, day, wind speed, wind direction, humidity, ignition source, and number of building floors) with high multicollinearity were excluded, and the remaining seven factors (month, hour, temperature, fire location, facility classification, ignition condition, and structural material) were selected to assess their importance using machine learning techniques.

In the second phase, SVM, RF, and GBRT were used, through which the validation and test data were used to calculate their MAE and RMSE. The results of this phase revealed that RF obtained the ideal results followed by GBRT and SVM. In the case of RF, at a tree number of 450, the difference between the MAE and RMSE values of the validation and the test data was the smallest; therefore, the overfitting problem was also the lowest.

In the third phase, as a result of analyzing the importance of the seven spatial and non-spatial fire factors using RF, the ignition condition was selected as the main factor in fire occurrence, followed in order by facility classification, fire location, hour, month, structural material, and temperature. The ignition condition factor includes carelessness and electrical and mechanical reasons. The analysis of fire data among these reasons demonstrates that fire accidents occurred frequently owing to carelessness, followed by electrical and mechanical reasons. On the basis of the above, as a result of applying the machine learning technique to the fire data in Seoul, it was possible to evaluate the important causes of fires, and the results are expected to be used as an important guideline for establishing urban fire management and reduction measures.

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 National Fire Data System: https://www.nfds.go.kr/index.do (accessed July 2021).
2 C. K. Lau, K. K. Lai, Y. P. Lee, and J. Du: Fire Saf. J. **78** (2015) 188. https://doi.org/10.1016/j.firesaf.2015.10.003
3 National Fire Protection Association: https://www.nfpa.org (accessed September 2021).
4 Y. Zhang: Procedia Eng. **52** (2013) 618. https://doi.org/10.1016/j.proeng.2013.02.195
5 J. Xin and C. Huang: Fire Saf. J. **62** (2013) 72. https://doi.org/10.1016/j.firesaf.2013.09.022
6 C. Hastie and R. Searle: Fire Saf. J. **84** (2016) 50. https://doi.org/10.1016/j.firesaf.2016.07.002
7 W. G. Song, H. P. Zhang, T. Chen, and W. C. Fan: Fire Saf. J. **38** (2003) 453. https://doi.org/10.1016/S0379-7112(02)00084-X
8 L. Telesca and W. Song: Chaos, Solitons Fractals **44** (2011) 558. https://doi.org/10.1016/j.chaos.2011.05.001
9 S. Lu, C. Liang, W. Song, and H. Zhang: Saf. **51** (2013) 209. https://doi.org/10.1016/j.ssci.2012.07.001
10 D. Rohde, J. Corcoran, and P Chhetri: Comput. Environ. Urban Syst. **34** (2010) 58. http://doi.org/10.1016/j.compenvurbsys.2009.09.001
11 V. Vapnik and A. Lerner: Autom. Remote Control **24** (1963) 774. https://cir.nii.ac.jp/crid/1571135650527018624
12 C. S. Yu, C. J. Lin, and J. K. Hwang: Protein Sci. **13** (2004) 1402. http://www.proteinscience.org/cgi/doi/10.1110/ps.03479604
13 L. Xing, W Yan, and J. Zhicheng: J. Syst. Simul. **33** (2021) 2606. https://doi.org/10.16182/j.issn1004731x.joss.21-FZ0705
14 L. Breiman: Mach. Learn. **45** (2001) 5. https://doi.org/10.1023/a:1010933404324
15 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
16 L. Breiman: Classification and Regression Trees, Chapman & Hall/CRC (Routledge., New York, 1984) 1st ed., pp. 59–218.
17 Z. Liu, G. Gilbert, J. M. Cepeda, A. O. Lysdahl, L. Piciullo, H. Hefre, and S. Lacasse: Geosci. Front. **12** (2021) 385. https://doi.org/10.1016/j.gsf.2020.04.014
18 T. W. Chu, A. Shirmohammadi, H. Montas, and A. Sadeghi: Am. Soc. Agric. Eng. **47** (2004) 1523. https://doi.org/10.13031/2013.17632
19 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
20 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
21 N. M. Abdaly, S. R. Taai, H. Imran, and M. Ibrahim: Enterp. Technol. **5** (2021) 59. https://doi.org/10.15587/1729-4061.2021.242986
22 E. J. M. Carranza and A. G. Laborte: Comput. Geosci. **74** (2015) 60. https://doi.org/10.1016/j.cageo.2014.10.004
23 M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan: Comput. Geosci. **148** (2021) 1. https://doi.org/10.1016/j.cageo.2021.104688

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. degree from Gyeongsang National University, Republic of Korea, in 2017. He is currently working on his Ph.D. degree. From 2000 to 2015, he worked on ICT and GIS at the Meteorology and Hydrology Service of Peru. His research interests include GeoAI 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 the Korean Society for Geospatial Information Science from 2009 to 2010. His research interests are in GIS and big data analysis.
(hhyoo@gnu.ac.kr)