MYU Tokyo

S & M 3149

Grid-based Urban Fire Prediction Using Extreme Gradient Boosting (XGBoost)

Haeng Yeol Oh and Meong-Hun Jeong*

Department of Civil Engineering, Chosun University,
309 Pilmun-daero, Dong-gu, Gwangju 61452, Republic of Korea

(Received October 24, 2022; accepted December 12, 2022)

Keywords: urban fire risk, grid-based prediction, machine learning, XGBoost

Fires in urban areas lead to enormous financial and human losses because cities have high densities of people and buildings. Although a recent advanced IoT technology improves early fire detection, it is crucial to predict fire risk to manage and prevent urban fires. We propose a method of predicting urban fires using extreme gradient boosting (XGBoost), which is based on grid-based data, to consider the characteristics of urban fires occurring in local areas. Before model training, we conducted a correlation analysis and variance inflation factor (VIF) analysis to remove variables with a strong correlation between independent variables. Furthermore, oversampling and feature selection techniques were applied to improve the model's performance. Experimental results revealed that the overall accuracy of XGBoost was 81.25%, the F1-score was 86.43%, and the area under the curve (AUC) was 84.59%. XGBoost performed better than baseline models, such as the support vector machine (SVM) and logistic regression. The results of this study show that it can be used for local area management and the prevention of urban fires.

1. Introduction

A high density of people and buildings can result in significant financial losses and many casualties in the event of a fire in a city. According to the International Association of Fire and Rescue Services (CTIF), the world's largest fire organization, 3082565 fires in 34 CITF member nations and 32 cities led to 87404 casualties in 2019.(1) In particular, fire-type data from 24 countries revealed that 812140 building and 344788 vehicle fires mainly occurred in urban areas, which account for 45% of the total 2572171 fires. Therefore, most fires occur at city centers, leading to the development of early IoT fire detection systems worldwide.

In the Republic of Korea, 47.4 million people reside in urban areas (about 91.8% of the total population as of 2021).(2) The total number of buildings in cities with more than 500000 people increased from 973297 in 2010 to 1,329,012 in 2021, an increase of approximately 36.54%.(3) As most of the population resides in urban areas and the number of buildings continues to increase, fire can cause significant damage to people and property at the city center. The Republic of

*Corresponding author: e-mail: mhjeong@chosun.ac.kr
https://doi.org/10.18494/SAM4194

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

Korea's Ministry of Interior and Safety (MOIS) reported in the 2020 Disaster Yearbook that in 2019 and 2020, there were respectively 10 and 7 cases of large-scale fires in multipurpose facilities, one of the types of social disasters.(4) Such a disaster is the most prevalent among the 28 types of social disasters defined by the MOIS. As a result, 79 people were injured, the second highest number of casualties among the various types of social disasters, and property damage was the third highest at 72.2 billion won. Therefore, it is necessary to prepare management strategies to address urban fires to move towards sustainable urban development.

Risk prediction studies for disaster preparation and prevention have been conducted in various fields, such as forest fire prediction(5–8) and crime prediction.(9,10) In urban fire prediction, Wang *et al.*(11) performed a grid-scale spatiotemporal prediction using the combining gate recurrent unit and conditional random field (GRU-CRF) model to predict fire risk in the Zhengzhou area. In this study, the relationships between urban data, such as education, medical care, and public facilities, were identified through correlation analysis. The CityGuard system, a web service for predicting urban fire risk, was implemented using the GRU-CRF model. However, in the case of correlation analysis, only the relationship with the dependent variable was analyzed. However, the correlation between independent variables was not analyzed, and fire-vulnerable building data that could lead to large fires were not used. Dang *et al.*(12) proposed a fire risk prediction model using publicly available data in the UK. Several types of fire-related data have been used for fire risk prediction, such as commercial property data in the Humberside area, fire reports, and safety status evaluation data. The experimental results revealed that among several machine learning algorithms, the AdaBoost algorithm was the most suitable fire risk prediction model. However, Dang *et al.* did not solve the problem of the class balance of fire categories in the fire data nor did they use granular data on buildings that were most closely related to financial damage in the event of a fire.

Currently, in the Republic of Korea, a spatial information system related to disasters, such as large-scale fires in multidensity facilities, is being built in administrative district units. Disaster management based on administrative district units can enable an easy evaluation of local governments or a quick compilation of statistics. However, because an area changes over time, it is not easy to compare the data on the same administrative district by dividing it by time. In addition, the analysis of patterns in nonuniform spaces is complex, and the analysis and statistics compilation of local areas are impossible. In particular, urban fires occur in local areas, not in administrative districts, so precise management and prevention are impossible at the administrative district level. Therefore, in this study, we present a grid-based fire occurrence prediction method for Suwon-si, Gyeonggi-do, Republic of Korea, that is similar to that in the study by Wang *et al.*(11) Furthermore, previous studies did not subdivide the fire-related building data, whereas we subdivided buildings into vulnerable buildings in the case of an urban fire in order to conduct urban fire risk prediction.

The remainder of this study is organized as follows. In Sect. 1, we discuss the importance of urban fire management and prevention using world statistics related to urban fires and the current situation in Korea. In Sect. 2, the data description and preprocessing for grid-based urban fire prediction are comprehensively explained. Section 3 contains an explanation of the extreme gradient boosting (XGBoost) and baseline models, and a description of how to evaluate these models. The results of the preanalysis and model evaluation for grid-based urban fire prediction are presented in Sect. 4. In Sect. 5, we present an overall summary and the limitations of this study.

2. Data

Various types of data can provide helpful information for urban fire prediction. However, these generally have different formats. We utilized a grid to integrate the data. Below, we describe the data and preprocessing process.

2.1 Study area

The target area of this study was Suwon-si, located in Gyeonggi-do, Republic of Korea (Fig. 1). Suwon-si is the 20th smallest out of 31 cities and counties in Gyeonggi-do. However, it has the largest population in Gyeonggi-do. Furthermore, because it has the highest percentage of residential facilities, food service businesses, and sales facilities in terms of buildings older than 30 years, it is more likely to be at a high risk of fire.(13) According to the 2017 fire statistics survey, 133 fires occurred in residential facilities belonging to high-risk groups in Seongnam City.(13) Suwon-si was considered suitable for this study because it corresponds to an area that is vulnerable to urban fires.

2.2 Data description

We collected nine datasets related to urban fires from Suwon-si in 2020 to predict urban fires. The National Fire Agency, the Ministry of Land, Infrastructure, and Transport, the

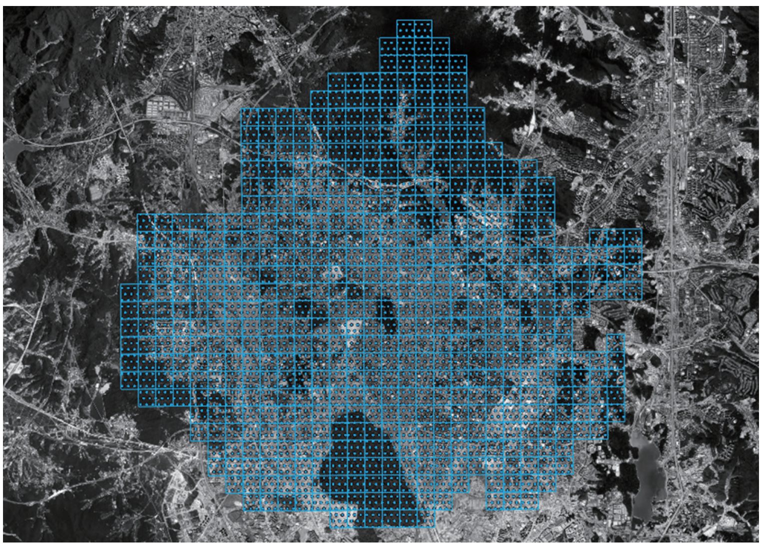


Fig. 1. (Color online) Suwon-si, located in Gyeonggi-do, Republic of Korea.

National Disaster Management Information System, and local governments collaborated to collect the data. A description of each data point is presented in Table 1. The number of building fires and population data were allocated to grids 500 m × 500 m in size. The entire area of the grids is consistent with that of Suwon-si, as shown in Fig. 1. The other data correspond to polygon data of buildings located in Suwon-si, which are building attributes.

2.2 Data preprocessing

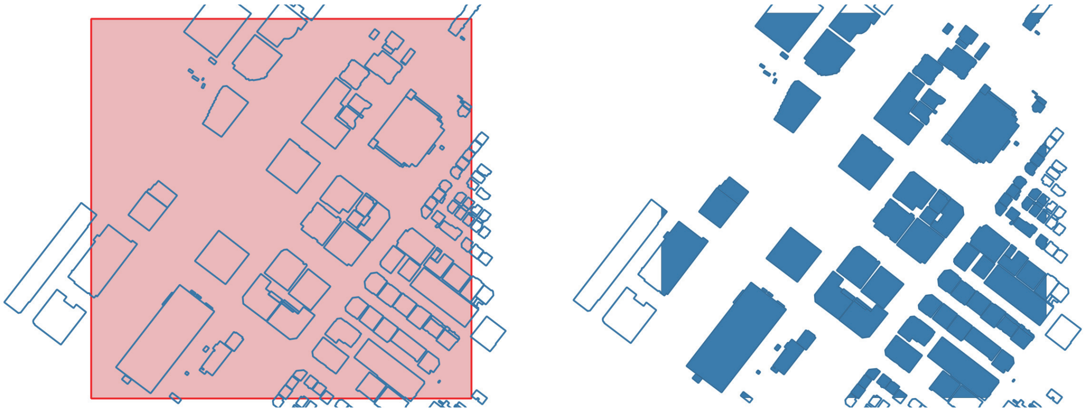
In this study, a grid was used as the basic unit of fire prediction. The spatial information should be converted into a grid. For example, the point data can be calculated by overlapping the number of points belonging to the grid. Polygon data must be divided using QGIS, a spatial information software program, in accordance with each grid size because they exist across the boundary of the grid, as shown in Fig. 2(a). QGIS can detect the intersection area between the polygon and the grid, as shown by the blue area in Fig. 2(b). Each area of the divided polygon data was then calculated and summed. The total area of the divided polygon data is then assigned to each grid.

The data used in this study had different units and ranges. Different data scales can be biased toward variables with large values because of the different distributions of the data, resulting in poor performance, and convergence can be hindered when using methodologies such as gradient descent. One strategy to solve this problem is normalization. Here, we utilized the commonly used normalization method, min-max normalization. Min-max normalization is a method of compressing or increasing data between 0 and 1, as shown by

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

Table 1
Description of data

Name	Description	Data type
Number of fires	Number of fires per grid	Grid
Population	Number of people per grid	Grid
Fire-vulnerable area	An area that the Fire Service deems necessary to designate as a fire boundary zone: 1. market; 2. dense factories/warehouses; 3. dense wooden buildings; 4. dangerous material storage and treatment facilities; 5. petrochemical complex; 6. industrial complexes; 7. fire boundary areas, designated by the Fire Department, without fire-fighting facility/water facility or fire-fighting route.	Polygon
High-rise building	A building with more than 30 floors	Polygon
Old building	A multidensity facility more than 30 years since completion (1990) as of 2020	Polygon
Multiuse facility	An area of a building that is used by an unspecified number of people and is likely to cause large-scale damage to human life and property in the event of an accident	Polygon
Fire-preventing area	A designated area between urban planning areas to reduce the risk of fire or other disasters in dense urban areas of the city center.	Polygon
Gas storage facility	A facility handling flammable materials, such as CNG filling, LPG filling, and gas stations	Polygon
Dryvit	A building whose exterior walls are constructed with dryvit	Polygon



(a) (b)

Fig. 2. (Color online) Polygon data preprocessing: (a) before and (b) after.

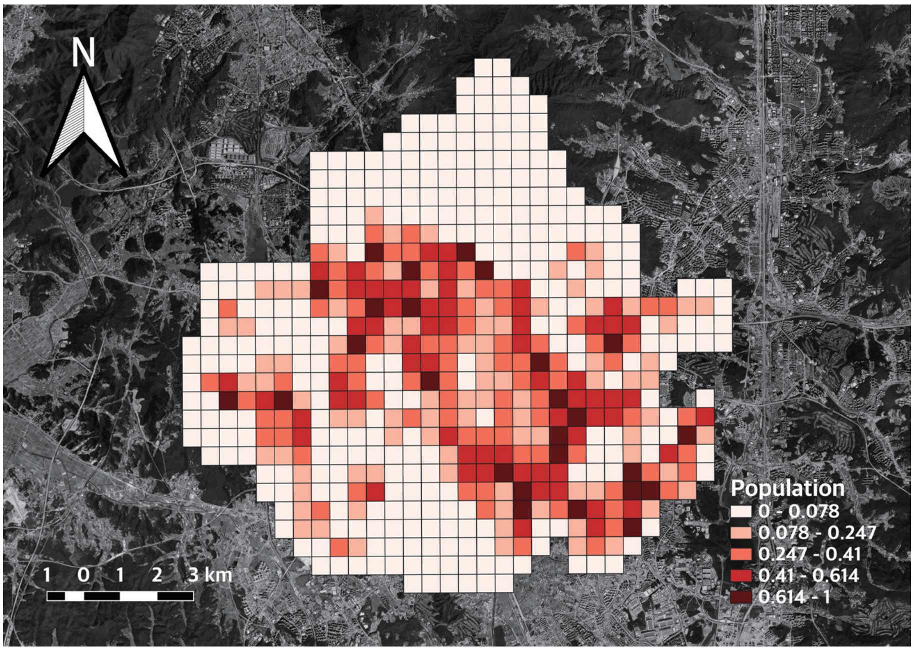


Fig. 3. (Color online) Visualization of the population grid in Suwon-si.

where x_{norm} is the normalized data, x is the original data, x_{min} is the minimum value of the data, and x_{max} is the maximum value of the data. Normalization was applied to all data except the number of fires in Table 1. Figure 3 shows an example of applying normalization to population grid data.

The data was divided for model training and evaluation. If a fire occurred in the grid, the value one was assigned to the grid and zero was assigned to the grid where no fire occurred. These values (i.e., 1 or 0) play the role of a dependent variable for predicting fire occurrence. We randomly assigned about 75% of data as training data and about 25% as evaluation data. Details are given in Table 2.

Table 2
Data separation

Data	Records	Non-fire grids	Fire grids	%
Train	447	332	115	74.24
Test	112	83	29	25.76

3. Methods

3.1 Models

We utilized XGBoost for fire prediction. XGBoost is a library implemented to support parallel learning with a gradient-boosting algorithm[14] using a combination of decision trees.[15] The advantage of existing tree-based learning is that the optimal split point can be determined by calculating the gradient for all cases of input data. However, analysis cannot be performed if all data are not loaded into memory. In addition, because gradients for all cases must be calculated, processing in a distributed environment is impossible, resulting in a low learning rate.

To solve the above problem, XGBoost uses an approximate algorithm. The approximate algorithm sorts the learning data for variable k, determines the percentiles in accordance with the distribution, and divides them by a certain number. The partitioned regions are called buckets, and the gradient for each bucket is calculated. Here, each bucket is grouped as either global or local, and epsilon is applied to each bucket. This approach reduces gradient computation while enabling parallel computation, leading to a faster and more accurate model training than with existing algorithms. Sparsity-aware split finding, column blocks for parallel learning, cache-aware access, block compression, and block sharding are also used to improve data processing and learning speed.

In this study, we also compare the performance of XGBoost with those of state-of-the-art alternative methods (i.e., support vector machines and logistic regression) for benchmarking. A support vector machine (SVM) is a supervised machine learning model. Classification and regression analysis are performed by constructing a hyperplane that maximizes the margin width between two categories in a high-dimensional space.[16,17] Here, the margin is the distance between the decision boundary separating the two categories and the support vector, which is the closest point. The optimal boundary is determined using the support vector, and the margin corresponds to the hyperplane. In an SVM, data that are difficult to separate into two dimensions are separated by mapping the low-dimensional space to the high-dimensional space using the kernel trick technique.

Logistic regression (LR) is a representative statistical technique used to predict the probability of the occurrence of a dependent variable using a linear combination of independent variables.[18] Unlike linear regression, LR is a classification technique in which the dependent variable targets categorical data. When input data are provided, the results are divided into specific categories.

3.2 Model evaluation

In this study, we used the overall accuracy (OA), F1-score, and area under the curve (AUC) to evaluate the urban fire prediction model.(19,20) The corresponding evaluation index was calculated using the confusion matrix presented in Table 3, and each index was calculated using Eqs. (2)–(5).

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \tag{2}$$

$$Recall = \frac{TP}{FN + TP} \tag{3}$$

$$Precision = \frac{TP}{FP + TP} \tag{4}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

AUC is the value of the area under the receiver operating characteristic (ROC) curve, and the model performance can be quantitatively evaluated from the ROC curve. The abscissa of the ROC curve is false positive rate (FPR) = $FP/(FP + TN)$ and the ordinate is true positive rate (TPR) = $TP/(TP + FN)$.

4. Results

Here, we present the results of correlation, the variance inflation factor (VIF), oversampling, and feature selection as the results of preliminary data analysis for precise model learning. Correlation analysis can discriminate variables that have a strong correlation with independent variables. VIF analysis can remove variables by identifying multicollinearity, which cannot be discriminated by correlation analysis. Furthermore, we use oversampling, which increases the amount of data when data are insufficient, thereby improving model accuracy. Finally, feature selection is applied to discriminate variables that are effective in model predictions among the oversampled data. The results of comparisons with the XGBoost model's hyperparameter optimization and model performance are then presented.

Table 3
Confusion matrix.

	Predicted fire grid	Predicted non-fire grid
Real fire grid	TP (true positive)	FN (false negative)
Real non-fire grid	FP (false positive)	TN (true negative)

4.1 Correlation analysis

Correlation analysis is necessary to train the model correctly. It is a method of analyzing the correlation between variables, where the correlation coefficient indicates the degree of linearity of the relationship between two variables. When a variable with a significant correlation coefficient between independent variables exists in machine learning models such as linear models or neural networks, the model's performance deteriorates or becomes unstable. Therefore, when there is a high correlation between independent variables, it is necessary to transform them into dimensions that are independent of each other or to remove one. We applied the Pearson correlation analysis. The experiment revealed no variables with strong correlation coefficients of 0.7 or more or −0.7 or less (Fig. 4).

4.2 VIF analysis

VIF analysis was used to determine whether an independent variable had multicollinearity.(19) Although the correlation analysis reveals the correlation between variables, there is no clear standard for removing variables. Multicollinearity is a phenomenon in which a strong correlation appears between independent variables. In general, if the result of the VIF analysis exceeds 10, it is judged that there is multicollinearity, and if it exceeds 5, it is considered noteworthy. The experimental results showed no multicollinearity, as the value of each variable was less than 10. Therefore, it is possible to use all values for fire prediction (Table 4).

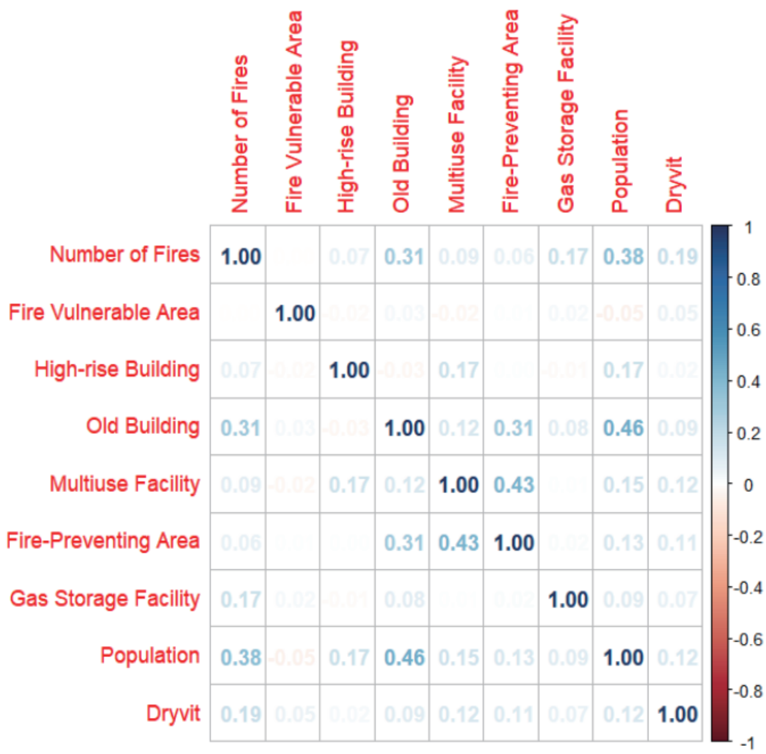


Fig. 4. (Color online) Results of Pearson correlation analysis.

Table 4
VIF analysis results.

Name	Fire vulnerable area	High-rise building	Old building	Multiuse facility	Fire-preventing area	Gas storage facility	Population	Dryvit
VIF	1.009297	1.074704	1.408019	1.285204	1.344133	1.014827	1.356660	1.036719

4.3 Oversampling

The training data from 447 records used in this study were imbalanced data consisting of 332 records of non-fire data and 115 records of fire data. If the model were trained using unbalanced data, it would be weighted to the category with the greater amount of data; therefore, training on the data of the category with less data may not be adequately performed.

One solution to this balance problem is to adopt a sampling technique that solves the imbalanced problem by increasing the number of categories with a small number of records or reducing the number of categories with a large number of records to generate better virtual samples. Sampling can be broadly classified into down-sampling and oversampling methods. Down-sampling is a method of removing category data with a large number of records; this can reduce learning time and prevent overfitting problems. However, important information for learning can be lost because of the removal of the original data. Therefore, down-sampling is appropriate when the number of records is sufficiently large.

The number of records of the training data in this study was minimal at 447, so it was inappropriate to apply down-sampling. Therefore, oversampling techniques such as the synthetic minority oversampling technique (SMOTE),(21) borderline-SMOTE (BLSMOTE),(22) density-based SMOTE (DBSMOTE),(23) and adaptive synthetic sampling (ADASYN)(24) were adopted. These methods do not cause information loss because they generate new data based on information from existing data. SMOTE is a method of generating data from categories with small amounts of data using the k-nearest-neighbors (KNN) algorithm. BLSMOTE is an algorithm that applies SMOTE only to samples corresponding to the boundaries of the data, judging that the borderline data greatly affect the class balance problem. DBSMOTE is an algorithm that creates a cluster using DBSCAN and then applies SMOTE within the cluster. ADASYN is a more advanced SMOTE algorithm that determines the amount of data to be generated in accordance with the distribution through the calculation of r, which is the density distribution of the data.

We applied the techniques mentioned above to XGBoost to select the optimal method. The accuracies obtained when using the various techniques are shown in Fig. 5. The use of original data resulted in an accuracy of 76.73%, SMOTE 79.54%, BLSMOTE 80.65%, DBSMOTE 80.59%, and ADASYN 82.48%. The ADASYN method performed better than the other methods. Accordingly, we used this method to solve the imbalanced data problem.

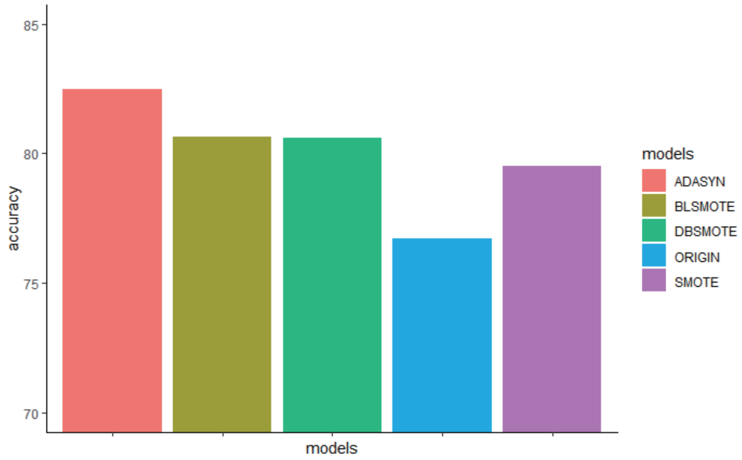


Fig. 5. (Color online) Performance of oversampling methods.

4.4 Feature selection

In addition to the correlation analysis between independent variables, it is essential to check which features are useful for prediction. We used recursive feature emission with cross-validation (RFECV), a feature selection technique, to filter out features unrelated to fire prediction.

RFECV is a method of finding the optimal variable for learning by calculating the feature importance of all variables, removing them from low-importance variables, and then calculating the performance of the model with the remaining variables. Random forest (RF) was utilized as a model for REFCV. The essential features were selected by K-fold cross-validation. When K was 10, six variables (fire vulnerable area, high-rise building, fire-preventing area, gas storage facility, and population) were selected. Figure 6 shows that the highest accuracy was 82.18% when these six variables were selected.

4.5 Model optimization

The XGBoost model was optimized in this study. The values of the hyperparameters were determined using a grid search and K-fold cross-validation. The K value was set to 10, and the optimal values of hyperparameters were determined with a grid search. Table 5 shows the list of the final hyperparameters for fire prediction after model optimization.

4.6 Model performance

The prediction performance characteristics of XGBoost, SVM, and LR are listed in Table 6. The OA of XGBoost is 81.25%, which is 6.25% higher than that of SVM and 5.36% higher than that of LR. Furthermore, the F1-score of XGBoost is 86.43%, which is 5.61% higher than that of SVM and 3.63% higher than that of LR. The AUC of XGBoost is 84.59%, which is 10.95%

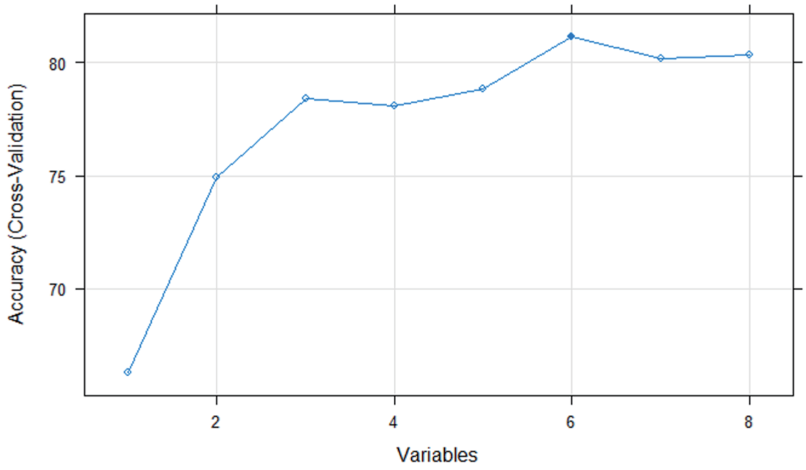


Fig. 6. (Color online) RFECV results

Table 5
Hyperparmeters.

Eta	Iterations	Max_depth	Max_leaves	Subsample	Sampling method
0.1	245	5	120	0.88	Uniform
Colsample bytree	Min child weight	Gamma	Lambda	Alpha	
0.96	8	0	4	2	

Table 6
Performance characteristics of different models.

Model	XGBoost (%)	SVM (%)	LR (%)
OA	81.25	75	75.89
F1-score	86.43	80.82	82.80
AUC	84.59	75.94	73.64

higher than that of SVM and 8.65% higher than that of LR. Therefore, XGBoost exhibits the best performance among the models considered.

5. Discussion and Conclusions

A precise analysis that considers various fire-related factors is required for the efficient management and prevention of urban fires. In this study, we predicted fires by transforming various types of data related to urban fires into a grid format.

Correlation and VIF analyses were performed before predicting fire to investigate whether there were strong correlations among independent variables. Oversampling and feature selection techniques were applied to improve the model's performance.

The experimental results demonstrated that the OA of XGBoost was 81.25%, the F1-score was 86.43%, and the AUC was 84.59%. XGBoost outperformed alternative methods such as SVM and LR. The findings of this study can aid in preventing urban fires and managing fire-vulnerable areas that are not identified in administrative district units.

However, the changes over time were not reflected in the current study owing to the limited amount of data. Therefore, future research should focus on an analysis that reflects the temporal and spatial factors of urban fires. In addition, rapidly developing deep-learning techniques can be applied to conduct fire predictions considering space and time.

Acknowledgments

This study was supported by a research fund from Chosun University, 2021.

References

1 World Fire Statistics: https://ctif.org/sites/default/files/2021-06/CTIF_Report26_KOREA.pdf (accessed August 2022).
2 LX Corporation: https://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=1200 (accessed August 2022).
3 KOSIS: https://kosis.kr/statHtml/statHtml.do?orgId=116&tblId=DT_MLTM_5391&conn_path=I2 (accessed August 2022).
4 MOIS: https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do;jsessionid=S6OLepLuc-2k5BZUsGkjYiqS.node30?bbsId=BBSMSTR_000000000014&nttId=89259 (accessed August 2022).
5 B. S. Negara, R. Kurniawan, M. Z. A Nazri, S. N. H. S. Abdullah, R. W. Saputra, and A. Ismanto: J. Phys. Conf. Series **1566** (2020) 012002. https://doi.org/10.1088/1742-6596/1566/1/012002
6 M. Naderpour, H. M. Rizeei, and F. Ramezani: Remote Sens. **13** (2021) 2513. https://doi.org/10.3390/rs13132513
7 B. Lee, S. Cho, S. K. Lee, C. Woo, and J. Park: Forests **10** (2019) 219. https://doi.org/10.3390/f10030219
8 F. Zhou, H. Pan, Z.Gao, X. Huang, G. Qian, Y. Zhu, and F. Xiao: Math. Problems Eng. **2021** (2021). https://doi.org/10.1155/2021/1929137
9 Y. L. Lin, M. F. Yen, and L. C. Yu: ISPRS Int. J. Geo-Inf. **7** (2018) 298. https://doi.org/10.3390/ijgi7080298
10 C. Huang, J. Zhang, Y. Zheng, and N. V. Chawla: Proc. ACM Int. Conf. Information and Knowledge Management (ACM, 2018) 1423–1432. https://doi.org/10.1145/3269206.3271793
11 Q. Wang, J. Zhang, B. Guo, Z. Hao, Y. Zhou, J. Sun, Z. Yu, and Y. Zheng: Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (ACM, 2019) 1–21. https://doi.org/10.1145/3369814
12 T. T. Dang, Y. Cheng, J. Mann, K. Hawick, and Q. Li: Proc. 25th Int. Conf. Automation and Computing (IEEE, 2019) 1–6. https://doi.org/10.23919/IConAC.2019.8894971
13 Institute, N. D. M. R.: https://www.ndmi.go.kr/research/research/view.jsp?DOC_ID=D0000018328 (accessed August 2022).
14 J. H. Friedman: Ann. Stat. **29** (2001) 1189. https://doi.org/10.1214/aos/1013203451
15 T. Chen and C. Guestrin: Proc. 22nd ACM Sigkdd Int. Conf. Knowledge Discovery and Data Mining (ACM, 2016) 785–794. https://doi.org/10.1145/2939672.2939785
16 C. Cortes and V. Vapnik: Mach. Learn. **20** (1995) 273. https://doi.org/10.1007/BF00994018
17 V. Arora, S. S. Rao, E. Amminedu, and P. J. Rao: Spatial Inf. Res. **29** (2021) 389. https://doi.org/10.1007/s41324-020-00359-1
18 D. R. Cox: J. R. Stat. Soc. **20** (1958) 215. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
19 A. Field, J. Miles, and Z. Field: Int. Stat. Rev. (2012) 276. https://doi.org/10.1111/insr.12011_21
20 H. Y. Oh, M.-H. Jeong, S. B. Jeon, T. Y. Lee, G. Kim, and M. K. Youm: J. Coastal Res. **114** (2021) 424. https://doi.org/10.2112/JCR-SI114-086.1
21 N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer: J. Artif. Intell. Res. **16** (2002) 321. https://doi.org/10.1613/jair.953
22 H. Han, W. Y. Wang, and B. H. Mao: Proc. Int. Conf. Intelligent Computing (Springer, 2005) 878–887. https://doi.org/10.1007/11538059_91
23 C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap: Appl. Intell. **36** (2012) 664. https://doi.org/10.1007/s10489-011-0287-y
24 H. He, Y. Bai, E. A. Garcia, and S. Li: Proc. IEEE Int. Joint Conf. Neural Networks (IEEE, 2008) 1322–1328. https://doi.org/10.1109/IJCNN.2008.463396