Article

Development of a Data-Based Machine Learning Model for Classifying and Predicting Property Damage Caused by Fire

Jongho Lee [1,2]**, Jiuk Shin** [3,*]**, Jaewook Lee** [1] **, Chorong Park** [2] **and Dongwook Sohn** [2]

[1] Korea Institute of Civil Engineering & Building Technology, Goyang 10223, Republic of Korea; leejongho@kict.re.kr (J.L.); juklee531@gmail.com (J.L.)
[2] Department of Architecture & Architectural Engineering, Yonsei University, Seoul 03722, Republic of Korea; sohndw@yonsei.ac.kr (D.S.)
[3] Department of Architectural Engineering, Gyeongsang National University, Jinju 52828, Republic of Korea
[*] Correspondence: jiukshin@gnu.ac.kr

Abstract: Large fires in factories cause severe human casualties and property damage. Thus, preparing more economical and efficient management strategies for fire prevention can significantly improve fire safety. This study deals with property damage grade prediction by fire based on simplified building information. This paper's primary objective is to propose and verify a framework for predicting the scale of property damage caused by fire using machine learning (ML). Korean public datasets are collected and preprocessed, and ML algorithms are trained with only 15 input data using building register and fire scenario information. Four models (artificial neural network (ANN), decision tree (DT), k-nearest neighbor (KNN), and random forest (RF)) are used for ML. The RF model is the most suitable for this study, with recall and precision of 74.2% and 73.8%, respectively. Structure, floor, causes, and total floor area are the critical factors that govern the fire size. This study proposes a novel approach by utilizing ML models to accurately and rapidly predict the size of fire damage based on basic building information. By analyzing domestic fire incident data and creating fire scenarios, a similar ML model can be developed.

Keywords: fire occurrence; property damage prediction; machine learning; data based; disaster prevention

1. Introduction

Over the past few centuries, cities have grown economically, scientifically, administratively, and culturally. As the cities have been urbanized quickly since the 20th century, the infrastructure of large cities has been built considering safety and security [1]. This is essential for minimizing damage in the event of natural as well as man-made disasters. At the same time, many studies especially on fire have been conducted to prevent and mitigate fires by constructing road networks for fire departments, emergency services, and disaster services [2]. In spite of the effort, issues related to urban fires, human casualties, and property damage have continued to occur worldwide since urbanization has made cities denser and more crowded, and a high population density has resulted in structures that are extremely vulnerable to disasters. For instance, in densely populated urban areas such as Seoul, even small disasters may cause severe damage [3]. Specifically, if problems, such as difficulty in the entry of fire trucks into the fire area, are not identified in advance, they may create a difficult situation for rescue and relief activities in the event of a disaster [4]. Therefore, social disasters such as fires have a high possibility of danger, and the failure of their initial suppression may result in large fires and massive damage, such as human casualties, owing to the rapid spread of fires to adjacent buildings.

According to the World Fire Statistics in 2019, the United States ranked first among 34 countries with 37,272,000 fires, with 77.5% of all fires occurring in residential areas. The problem is that accurately identifying the number of fires in the United States is difficult

due to the frequent occurrence of disasters, large fires, and terrorism. To be specific, urban fires have been identified as a serious problem, as can be noted by the death of 492 people in a fire at the Cocoanut Grove nightclub in Boston in the 1940s and 61 people in a fire at the LaSalle hotel in Chicago.

Frequent fire incidents in urban factories in South Korea have caused severe human casualties and property damage. According to a report released by the National Fire Agency (NFA) of Korea, 12,645 factory fires occurred over five years from 2016 to 2020, resulting in 900 casualties (70 deaths and 830 injuries). In the fire investigation and reporting regulations of South Korea, large fires are defined as the fires that cause more than five deaths or ten casualties or those that cause property damage of more than five billion KRW. For example, two fire incidents that broke out in factories have been classified as large fires, such as the fire at an electronics factory on August 21, 2018 that caused nine deaths and six injuries, and another at a logistics center in 2020 that resulted in 38 deaths and 12 injuries. As reported in the media, fires have frequently occurred in domestic industrial complexes because of the spread of fire caused by their dense structure and several explosions caused by chemicals. Furthermore, these have been reported to cause secondary damage to surrounding areas.

Regarding factory fires that cause numerous human casualties and property damage, damage can be minimized through various means, such as early fire suppression, occupant evacuation plans, and securing the stability of structures. Above all things, fire safety can be significantly improved by preparing more economical and efficient management strategies for fire prevention. Specifically, predicting property damage caused by fire is fairly difficult. For example, considerable time and manpower are required to secure data [5].

The research question of this study is as follows: can the degree of property damage in the event of a factory fire be predicted through a machine learning model learned using simple data such as building register information and fire scenarios? The size of the fire, the size of combustibles in the building, quick detection, and first responders have a significant impact on the size of the fire. However, this information is impossible to obtain before a fire occurs. However, this study begins with the hypothesis that when building information and fire scenarios for numerous fire events are learned, excluding information that cannot be secured, trends in fire size will be evident. Thus, this paper's primary objective is to propose and verify a framework for predicting the property damage scale caused by fire using machine learning (ML) with simple data. Korean public datasets are collected and used as training data for ML algorithms, and the accuracy of the proposed method is verified.

The remainder of this paper is organized as follows. Section 2 presents the trend of previous ML research related to fire to set the direction of ML development. Section 3 presents the material and methods (preprocessing of learning data and developing ML models). Section 4 presents results and conclusions derived from the development of a fire data-based property damage rating classification model. Finally, Section 5 presents the conclusions.

2. Literature Review

To set the research direction for developing an ML algorithm for predicting the property damage scale, which is the purpose of this study, we examine the latest trends in artificial intelligence (AI) research in the field of fire. As a result, the papers with keywords, such as fire + AI, were examined, and they were classified according to fields, such as architecture, civil engineering, and firefighting. The input data used in each study included material performance, images or videos, environmental information (e.g., temperature and humidity), and fire causes. Fire occurrence was mostly predicted as output data; however, certain studies predicted the damage scale, human casualties, and fire stages in more detail. The details are presented in the subsequent subsections.

2.1. Research Trends in Architecture

Generally, the contents of the papers were related to data utilization for collecting or predicting AI-based fire information. Based on this, research was mainly conducted to predict the risk of architectural structures vulnerable to fire [6]. To be specific, studies on data utilization have constructed a smart framework that can predict smoke movement and availability in the event of a fire in a building. Further, the framework research has been developed for the construction of algorithms that can secure time for safe evacuation (available safe egress time; ASET). In the process of the construction of algorithms, smoke movement was predicted by constructing a database by preparing profiles, such as the length of the atrium that can perform ventilation, fire size, ventilation conditions, and time after ignition [7]. In addition, other related studies have conducted the real-time prediction of temporary fire scenarios using external smoke images and deep learning algorithms. A comprehensive collection of 1845 large-scale databases was formed, showing that hidden fire information can be determined in real time by training deep learning algorithms with the smoke images simulated using convolutional neural network (CNN) models. The large potential of smart firefighting was also demonstrated. This verified the possibility of using AI for firefighting performance-based design (PBD), which can reduce the time and cost required to create a fire safety architectural environment [8].

In addition, reducing fire damage by deriving elements vulnerable to fire in architectural structures based on the developed AI was considered. An AI-based cognitive framework was constructed to track the reactions of concrete structures among building structures to high temperatures. This algorithm can successfully understand the natural and complex behavior of reinforced concrete (RC) structural members exposed to fire. Furthermore, it considers the characteristics of concrete and steel reinforcement at high temperatures and related phenomena [9]. In other studies, a series of ML models were constructed to predict the fire risks of buildings based on their structural characteristics. Approximately two-thirds of fires that occurred at research sites were accurately classified through learning with data, including the structures of fire buildings and building-level information. These algorithms are expected to help reduce fire damage by excluding uncertain factors and utilizing data that can be objectively measured in data analysis for building fire prediction.

2.2. Research Trends in Civil Engineering

In research related to civil engineering, studies were conducted for providing scientific guidelines for smart firefighting technology and future emergency response tactics in smart cities by outputting the characteristics of structures that can spread fire risks using AI. In terms of the structure of buildings, research was conducted to construct a prediction model that used the results of the conducted numerical analysis as input data and could generate the fireproof output of the RC columns embedded in walls for the given input data [10]. In particular, the element geometric effect, concrete cover thickness, reinforcement ratio, axial strength, and bending moment were analyzed as dominant factors that affected the fire resistance of eccentrically loaded columns, which are part of the fire prevention room walls of buildings. Further, a prognostic model capable of generating output for the fire resistance of this type of RC column was constructed for the given input data, and fire resistance curves were derived based on the results obtained through numerical analysis and the neural network prognostic model [11]. Subsequently, the fire risk of tunnels, which are one of the road facilities, was predicted. The fire causes were predicted in numerical models for tunnels via the application of AI and big data framework, and a large-scale tunnel fire database of numerical simulation with various fire locations, fire sizes, and ventilation conditions was constructed. The temperatures measured using various sensor devices were used to train long-term memory recurrent neural networks, and it was found that the location, size, and ventilation wind speed of a tunnel fire could be predicted with 90% accuracy when using a trained model. These studies are expected to show the possibility of predicting fire causes and risks based on AI.

2.3. Research Trends in Firefighting

Firefighting research has generally focused on data collection and prediction to facilitate a swift response of fire authorities in the event of a fire [12]. First, for accuracy in fire prediction, research was conducted on new algorithms to predict fire risks for properties based on ML. The data required for statistical learning must be composed of numerical data, such as variables that affect fire occurrence and reaction variables that indicate the fire frequency. Algorithms have been implemented through statistical ML for the numerical data. Further, ML and deep learning have been used, including several datasets, such as fire management, effective response to fire, fire spread prediction, and detection. It was proven that implementing algorithms that use other frameworks' data can accurately predict fire occurrence [13]. In addition to statistical ML, FireCast, a new system that combines AI technology with geographic information systems (GIS) data collection strategies, can predict the areas around burning forest fires prone to high fire risks in the foreseeable future. FireCast outperformed the random prediction model and Farsie, a commonly used forest fire diffusion model, in terms of total accuracy, recall, and F-score [14]. Further, studies have focused on fire prediction systems through image recognition and prediction methods through ML by constructing numerical datasets. Multi-sensor detection systems were combined with image recognition to implement rapid and stable smart fire detection systems, and research was conducted to extract and classify important features from the existing images collected in real environments using ML. Studies have also attempted to identify the structure of buildings in the event of a fire to utilize the prediction systems of fire authorities [15]. An FE-based ML framework was developed to predict structural response to fire in real time based on the temperature data of structural members, and a numerical database was constructed for steel structures that are affected by hundreds of fire scenarios. Structural response to fire was simulated in ABAQUS using the FE method. The FE-based ML framework developed in the study can predict the real-time response of a structure to fire using the ML model based on the FE database. This verified that it can supplement the considerable time consumption of the traditional FE method when applied to a fire emergency. These studies are expected to aid officials and fire authorities in managing resources more efficiently. Further, they can facilitate the prevention of disasters by proposing optimized models focused on risks using statistical ML and indexing for fire risk assessment.

Large fires further increase human casualties and property damage, and studies on classifying previous fires through data analysis that applied ML and predicting fires in advance have been conducted to prevent them. Numerous countries have constructed fire databases that can be used to predict and manage fires. AI-based research, which combines various types of data (e.g., images, videos, and big data) with fire data and performs training using AI, can improve fire detection and prediction accuracy. This can aid in minimizing the risks or damage of fires that may occur in urban spaces in the future.

The study advances the previous literature by proposing and verifying a framework for predicting the property damage scale caused by fire using machine learning (ML). While previous studies have focused on predicting the occurrence of fires, this study specifically addresses the prediction of property damage ratings based on simple building information. As this study aimed to develop a framework for prior response by predicting fire damage using the relatively simple and numerical information of buildings, it is significant because it enables rapid and swift fire damage prediction.

This study contributes to the previous literature on the research topic by addressing the prediction of property damage ratings based on simple building information, a relatively unexplored area in previous studies that have focused on predicting the occurrence of fires.

3. Material and Methods

This study aimed to develop an ML model that predicts the degree of damage in the event of a fire in a factory by learning ten-year factory fire data.

The methodology is shown in Figure 1. Preprocessing of data is required for ML. The steps of data preprocessing are as follows: (1) identifying and removing outliers, (2) selecting factory building fires, (3) selecting building in use, i.e., the fire that occurs under construction or in the process of demolition can have a negative impact on the accuracy of the results, (4) excluding small fires which are less than 1 m^2, i.e., very small fires are difficult to be analyzed precisely, (5) adjusting the levels of categorical variables, i.e., this step represents recategorizing, (6) changing nominal data to numeric data, i.e., this stage is needed to encourage learning efficiency, (7) generating derivative variables, and this step creates more meaningful variables, and (8) setting the dependent variable.
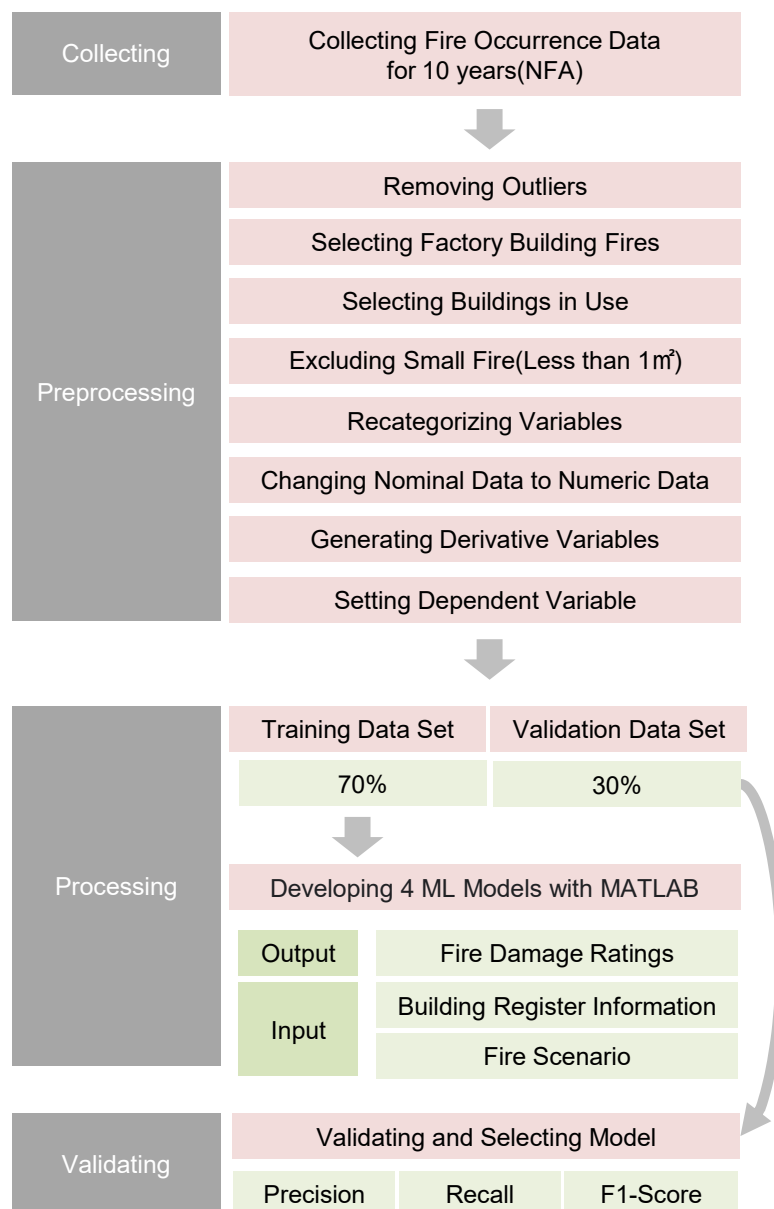
Collecting	Collecting Fire Occurrence Data for 10 years(NFA)

	Removing Outliers
	Selecting Factory Building Fires
	Selecting Buildings in Use
Preprocessing	Excluding Small Fire(Less than 1㎡)
	Recategorizing Variables
	Changing Nominal Data to Numeric Data
	Generating Derivative Variables
	Setting Dependent Variable

	Training Data Set	Validation Data Set
	70%	30%
Processing	Developing 4 ML Models with MATLAB	
	Output	Fire Damage Ratings
	Input	Building Register Information
		Fire Scenario

	Validating and Selecting Model		
Validating	Precision	Recall	F1-Score

Figure 1. Methodology and conceptual workflow.

After data preprocessing, learning of four different models by MATLAB is carried out. Out of the dataset, randomly picked 70% and 30% of data are used for learning and verifying, respectively. Then, the learning process is implemented twice to check if the performance is ensured with only simpler information. The first model merely learns the building of register information and the second model learns fire scenarios in addition to

the first model. Finally, the model with the greatest performance is selected, examining precision, recall, and f1-score of the four models developed with 30% of the dataset.

3.1. Dataset Construction

The dataset used in this study was the national public data collected by NFA. The dataset is highly reliable as a national agency has constructed it. Fire size is typically influenced by factors such as combustibles within the building and firefighting equipment (such as sprinklers). But it is difficult to obtain such information unless one is a building owner or manager. This study aims to predict fire size using relatively simple data that are publicly available; therefore, it does not include specific information on the building. The utilized data were collected over ten years, from 2009 to 2018. During this period, a total of 433,737 fires occurred in Korea. Here, the entire data was analyzed only for building fires, excluding forest, automobile, railway, aircraft, and ship fires. In addition, among the buildings, only factories in operation were considered. Based on the fire growth theory [16], a burnt area of 1 m^2 or less was defined as a small fire and excluded from the data. A burnt area of 1 m^2 or lower implies a fire wherein the ignited local area was burnt. Thus, fire damage was not likely to increase. Consequently, 12,223 items (rows) were filtered, and the total number of variables used in the analysis was 16 (columns).

3.2. Variable Information and Data Preprocessing

The data provided by NFA consists of continuous and categorical types. As listed in Table 1, continuous variables were divided into seven types, including property damage, number of floors, number of basement floors, total floor area (TFA), and building area. Categorical variables were divided into eight types, including the ignition heat source and ignition factor. Among the continuous variables, one target variable (property damage) was set, and those without property damage were used as predictive variables in the analysis.

Table 1. List of data variables.

Category	Variable	Number
Continuous type	Property damage, number of floors, number of basement floors, TFA, building area, burnt area, and number of casualties	7
Categorical type	Facility location information, structure, industry type, month of fire, time of fire, ignition factor, ignition material, and ignition point	8

For the categorical variables, the model performance decreased when the frequency (or ratio) of each variable level was low. Thus, recategorization was performed by integrating and modifying the number of levels of the categorical variables. Recategorization removes outliers and refers to reducing the number of levels (classes) of the categorical variables. It is considered when one categorical variable has ten or more levels. Rare events that are difficult to occur or levels with a low frequency are eliminated, and the number is adjusted to approximately four to improve the performance of the classification model. Recategorization was performed for the final 12,223 data items. Twelve months were recategorized into four seasons for the month of fire and 24 h into four 6 h sections for the time of fire. For facility location information, 226 locations were recategorized into the metropolitan management area (MMA), metropolitan cities (MCs), and provinces/regions (PRs). This is shown in Table 2.

Table 2. List of variables to be recategorized.

Recategorization Sequence	Target Variable	Number of Recategorized Levels	
		Before	**After**
1	Month of fire	12 (Month)	4 (Season)
2	Time of fire	24 (Hours)	4 (6 h)
3	Facility location information	226 (District)	3 (Region Area)

3.3. Derivative Variable Generation and Key Variable Selection

Derivative variables were used to improve the reliability and accuracy of the classification model, and new variables were generated based on the existing variables to discover significant factors for the model. Generally, they are generated by combining individual variables at a commonsense level. For example, they were generated by applying the four fundamental arithmetic operations on continuous variables and logical values between variables (e.g., whether certain conditions are applied) for the categorical variables.

As shown in Table 3, derivative variables were generated based on TFA, burnt area, number of casualties, and property damage, which are continuous variables. According to the Building Act of Korea, a fire-resistant structure is mandatory for a TFA of 5000 m^2 or larger. Thus, it was set as a standard.

Table 3. List of generated derivative variables.

Order	Name of the Derivative Variable	Variable Type	Number of Levels
1	Fire-resistant structure	Categorical	2
2	Burnt area/TFA	Continuous	-
3	Property damage rating	Categorical	3
4	Human casualties	Categorical	2

In particular, the fire damage rating was set as the dependent variable to increase the learning success rate. Following the classification of the property damage, the distribution of each rating was analyzed and adjusted to have a distribution rate of approximately 33 (Table 4).

Table 4. Fire damage ratings and dependent variable setting.

Fire Damage Rating	Range (KRW)
Small scale (low)	$\leq$3,600,000
Middle scale (moderate)	>3,600,000 $\leq$25,000,000
Large scale (severe)	>25,000,000

The variables derived through the aforementioned process are shown in Table 5. The proposed ML is developed in Section 4, and its performance was examined using only building register information. Further, the fire scenario information was added to the building register information to compare the learning model's performance.

Table 5. List of variables used in the classification model for the property damage caused by fire.

Number	Variable Name	Feature	Variable Type	Type of Use
1	Property damage rating	Output	Continuous	Independent
2	Facility location information	Building register information	Continuous	Independent
3	Number of floors	Building register information	Continuous	Independent
4	Number of basement floors	Building register information	Continuous	Independent
5	TFA	Building register information	Continuous	Independent
6	Building area	Building register information	Continuous	Independent
7	Structure	Building register information	Categorical	Independent
8	Industry type	Building register information	Categorical	Independent
9	Fire-resistant structure	Building register information	Categorical	Independent
10	Season of fire	Fire scenario	Categorical	Independent
11	Time of fire	Fire scenario	Categorical	Independent
12	Ignition factor	Fire scenario	Categorical	Independent
13	Ignition material classification	Fire scenario	Categorical	Independent
14	Ignition point classification	Fire scenario	Categorical	Independent
15	Human casualties	Fire scenario	Categorical	Independent
16	Burnt area/TFA	Fire scenario	Continuous	Independent

This was undertaken to examine whether a prediction is possible with only the data provided by the national agency first because the building register information is the only national data that can be obtained for fire damage prediction, and researchers must set certain values in fire scenarios for prediction.

3.4. ML Classifier Model Overview

In this study, four machine learning models were used. First, the artificial neural network (ANN) model is an ML methodology that describes the learning process of the human brain using mathematical and probabilistic methodologies. It comprises input, hidden, and output layers. Each layer has multiple nodes, and each node is combined with one or more other nodes. The nodes are connected complexly, delivering information on weight and bias values. The activation function converts the sum of the weight values into an output signal and transmits the related information to the next layer [17]. Examples that use this technique can be found in [18–21].

The second model is the decision tree (DT) model, an AI algorithm implemented with a tree-shaped model. It learns patterns existing between data by analyzing the data and estimates results by combining them. It performs learning by forming a tree structure from upper to lower nodes and selecting classification variables and criteria for each stage. The depth can be considered a representative hyper-parameter; however, overfitting is highly likely to occur in the training dataset with increased depth [22]. Examples that use this technique can be found in [23–26].

The third model is k-nearest neighbor (KNN). The KNN algorithm examines ambient data and classifies them into many data-containing categories. It is used assuming that data with similar characteristics can be included in similar categories. The algorithm's performance changes significantly depending on the k value, which indicates the number of ambient data. The k value has the most significant impact on learning performance [27]. Overfitting occurs with an increased k value because it is difficult to clearly express the features of the data. In contrast, overfitting may occur with the decrease in the k value under the influence of certain data [28]. Examples that use this technique can be found in [29–31].

The final model is the random forest (RF) model. Ensemble learning is the method of learning data using multiple learning models rather than a single ML model, and representative methods include voting, bagging, boosting, and staking [32]. The RF methodology is included in the bagging method. It is an ensemble model for DT and a collective model that calculates results by combining multiple decision trees with different characteristics. It exhibits high accuracy and can be used as a solution to the overfitting issue found in the

DT method [33]. The bagging method of the RF methodology develops multiple DT-based classification models by constructing multiple sub-datasets in the same dataset, and results are then estimated based on these models [34]. Examples that use this technique can be found in [35–38].

Table 6 shows ML models' features. It summarizes the advantages and disadvantages of ML models that are commonly seen in many studies.

Table 6. Learning models' features (advantages and disadvantages).

ML Classifier Model	Advantage	Disadvantage
ANN	Excellent predictability	Limitations in interpreting the results
DT	Ease of interpreting the results	High probability of overfit
KNN	Error data does not affect results	More data significantly reduces learning speed
RF	Prevents overfitting for high accuracy even with a high percentage of missing values	Limitations in interpreting the results

MathWorks MATLAB r2023a (v.9.14.9.2239454) was used for implementing the above ML methodologies.

4. Results and Discussion

4.1. Development of ML Models Using Building Register Information

This study aimed to predict the fire size based on minimal data. In this subsection, learning was performed using only the building register information provided by the national agency among the fire data introduced in Section 3, and the precision and recall of fire damage prediction were examined. The abbreviations and ranges of building register information variables to predict fire damage are as follows.

- Facility location information: MMA, MCs, and PRs.
- Fire-resistant structure: Yes or No.
- Industry type: Metal Machinery and Equipment Industry (MMEI), Wood Processing and Carpentry Industry (WPCI), Chemical Industry (CI), Food Industry (FI), Textile Industry (TI), Electrical and Electronics Industry (EEI), Pulp and Paper Industry (PPI), etc. (other industries).
- Structure: Steel Frame Structure (SFC), Reinforced Concrete Structure (RCC), Sandwich Panel Structure (SPC), Block Structure (BLC), Container Structure (CC), Brick Structure (BC), etc. (wood, greenhouse pipe, stone, steel frame reinforced concrete, brick veneer, container, and other structures).
- Number of floors: $1 \leq 30$.
- Number of basement floors: $0 \leq -4$.
- TFA: $1 \leq A_{tf} \leq 69{,}437{,}392$ m^2, mean: 12,813.30 m^2, standard deviation: 686,213.77 m^2.
- Building area: $0.03 \leq A_{fa} \leq 51{,}226{,}276$ m^2, mean: 9366.09 m^2, standard deviation: 488,293.30 m^2.

The database had a total of 12,223 data items. Figure 2 shows the distribution and frequency of the independent variables described earlier.

A regression analysis was conducted on the fire damage ratings included in the input and output information to investigate probabilistic correlations between the input and output variables used for learning. In general, if the *p*-value of an input variable is less than 0.05, the input variable can be considered statistically important because it has a significant impact on the output variable.
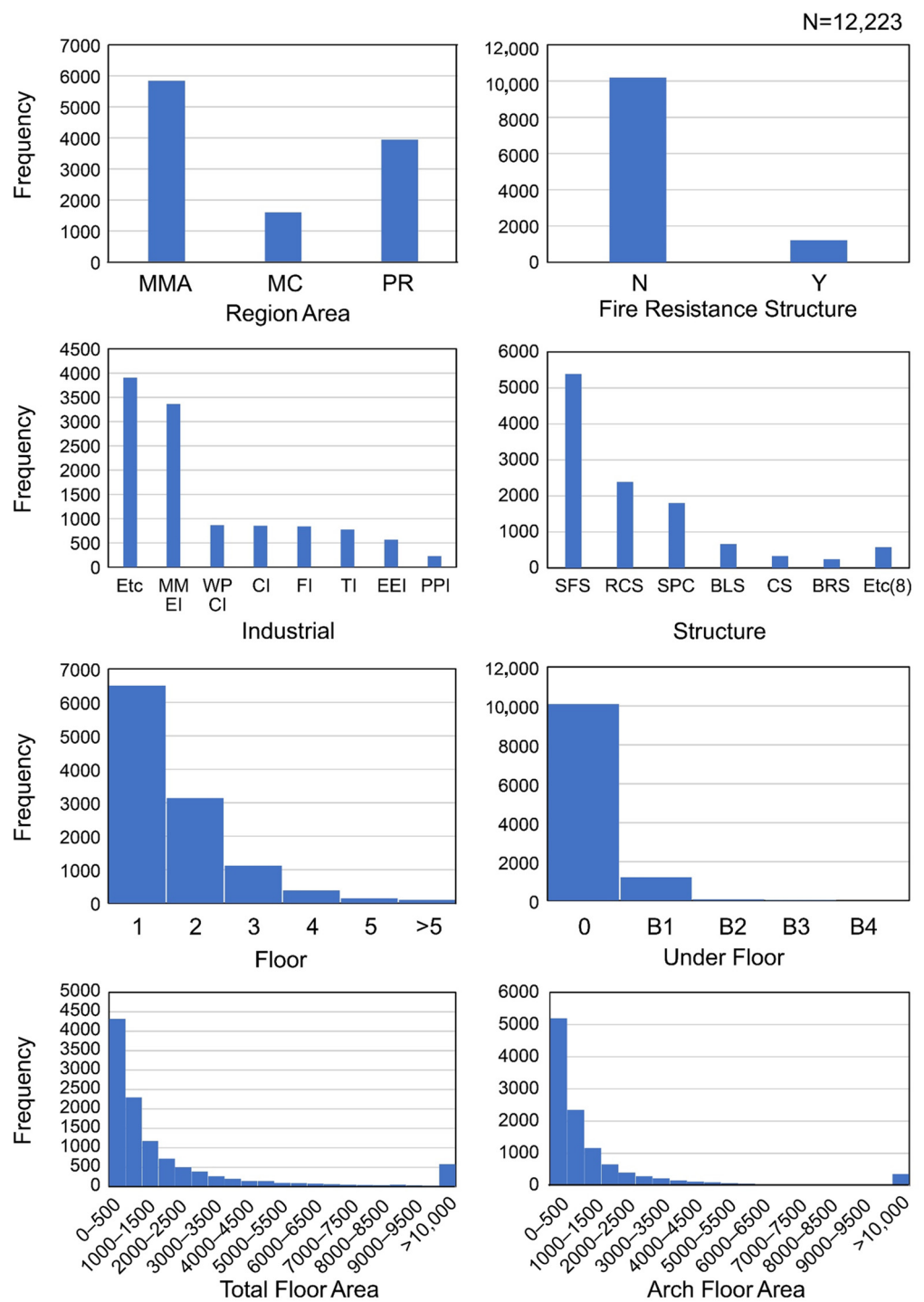
Figure 2. Distribution and frequency of building register information among fire data.

Table 7 shows that most variables are not statistically significant except the building location information, structure, industry type, and fire-resistant structure. Compared to other input variables, the building TFA and floor area variables were unimportant in determining the fire damage rating. This is because factory facilities with small TFA or floor areas cannot significantly affect the damage rating determined by the amount of damage. Thus, small factory facilities may not be considered important in damage rating classification because they can be calculated only as small damage ratings.

Table 7. Regression analysis on input variables for output variables (building register information).

Category		β	SE	F	*p*-Value
Building register information	Building location information	0.080	0.024	10.646	<0.001
	Number of floors	0.047	0.037	1.578	0.209
	Number of basement floors	−0.072	0.037	3.79	0.052
	TFA	−0.039	0.036	1.208	0.272
	Building area	−0.005	0.024	0.045	0.833
	Structure	0.184	0.031	35.642	<0.001
	Industry type	0.115	0.022	27.132	<0.001
	Fire-resistant structure	0.066	0.028	5.436	0.02

However, ML models were developed using all selected data without probabilistic judgments between the data.

The ML technologies described in the previous section were used to identify fire damage ratings from the constructed database. To use the ML models, in this study, eight variables out of the 16 variables, described in Section 3.3, were converted into input parameters. The ML code for the models, mentioned in Section 4.1, was developed using the Mathworks open source. Here, 70% of the 12,223 data items were used as training data (training set) and 30% as validation data (test set). The entire dataset was randomly divided into the training and test sets, and the model performance for the test set was an indicator that represented the model's performance for unknown data. In other words, the ML models were developed with 70% of the collected data using the methods mentioned in Section 4.1.

The performance of each ML model was evaluated in further detail using the confusion matrix (Figure 3). The figure shows the confusion matrix of the training and validation datasets used to compare actual and predicted values. The confusion matrix can compare the actual value (rating) for the given input variables with the value (rating) predicted by the ML model for the same input variables. The rows of the matrix indicate the predicted values, and its columns represent the actual values. As the values located in the diagonal cells (row 1, column 1; row 2, column 2; and row 3, column 3) show that the actual value (rating) and predicted value (rating) are identical, they indicate success in prediction through the ML models. The other cells in the matrix show that the rating was underestimated or overestimated. For example, the value placed in row 1, column 2 indicates that the ML model predicted a higher rating (moderate) than the actual rating (low) (overestimation), while the value in row 2, column 1 implies that the ML model predicted a lower rating than the actual rating (moderate) (underestimation).

It was found that the ANN optimization ML model developed earlier (classifier) accurately predicted the severe rating compared to other ratings. However, its success rate in predicting the moderate rating was low.

The DT optimization ML model (classifier) accurately predicted the severe rating (>80%); however, its success rate in predicting other ratings was significantly low.

The KNN optimization ML model (classifier) successfully predicted the severe rating (>50%) based on the validation dataset; however, its prediction success rate for other ratings was approximately 30%.

The RF optimization ML model (classifier) successfully predicted the severe rating (approximately 80%) based on the validation dataset; however, its prediction success rate for other ratings was less than 20%.

In addition, precision and recall were compared and analyzed to select a classifier learning model based on the values calculated from the confusion matrix. Precision is an indicator that represents the accuracy of prediction, while recall indicates the proportion of the data accurately predicted by the classifier and is an indicator that shows the classifier's sensitivity. Precision and recall have a trade-off relationship wherein an increase in either causes a decrease in the other [37].
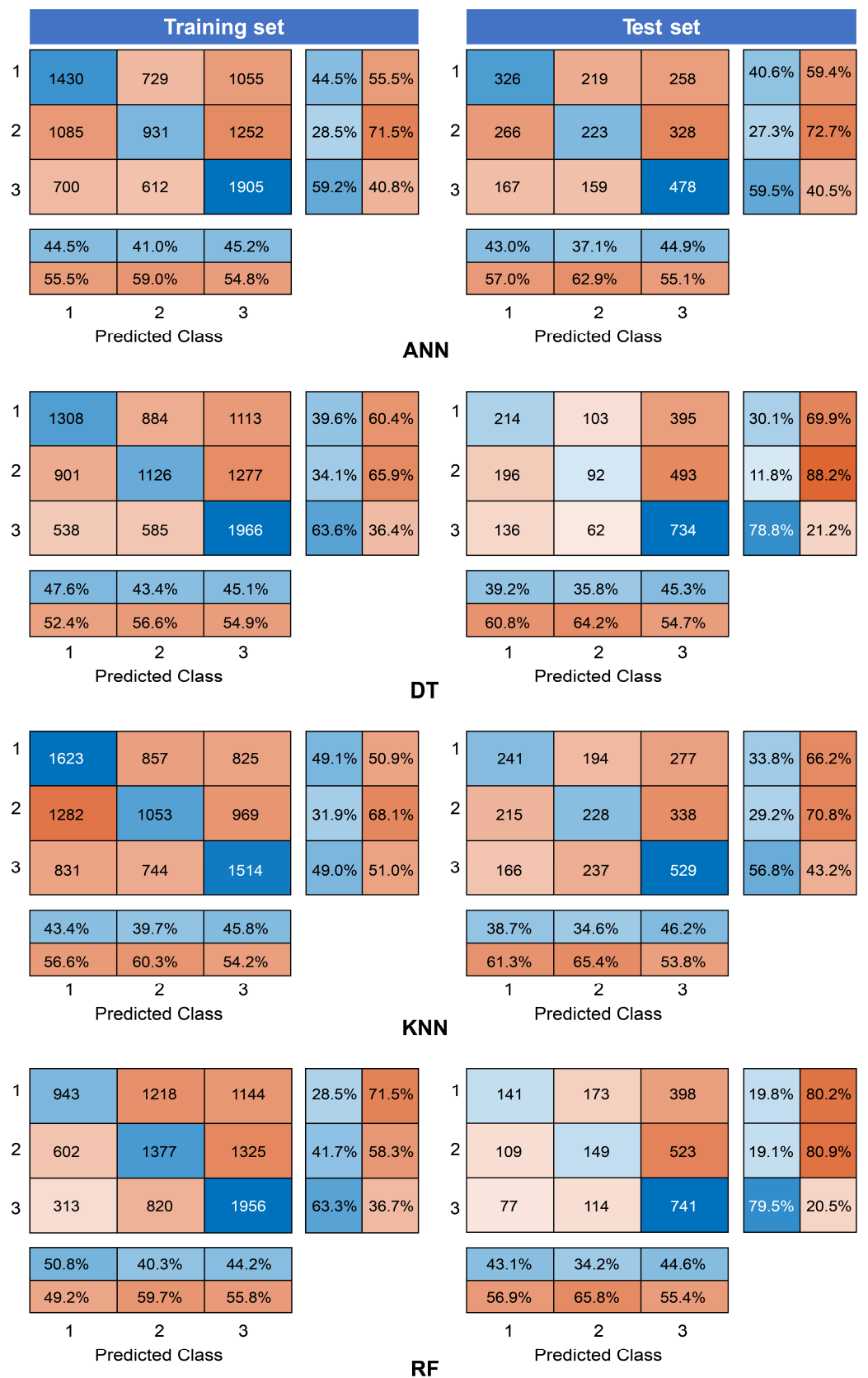
Figure 3. Confusion matrix of ML models (building register information).

For all four models, the precision and recall values for the moderate damage rating could not reach the average values, and those for the low and severe damage ratings were slightly higher than the average values.

Additionally, it was found that the recall and precision values were almost similar for all the four models. Since the models were developed for prior response by predicting fire damage using simple information, it is determined that classifiers with higher recall than precision are appropriate.

4.2. Development of ML Models Using Building Register Information and Fire Scenarios

As mentioned in Section 4.2, securing both precision and recall for estimating fire damage only with building information is challenging. Thus, fire scenario variables were used in addition to the data learned in Section 4.2 to develop ML models. As 830 data items out of the 12,223 data items had missing values for fire scenario data, they were excluded, and 11,393 data items were used for learning. As they corresponded to 6.8% of the total data and caused no significant change in distribution and frequency, the data analysis conducted in Section 3.4 was omitted.

The abbreviations and ranges of fire scenario variables to predict fire damage are as follows:

- Season: spg (spring), smr (summer), Fal (fall), and win (winter).
- Time of day: 06–12, 12–18, 18–24, and 00–06.
- Human casualties: Yes or No ($1 \leq$).
- Burnt area/TFA: $0.0 \leq A_{tf}/A_{fd} \leq 1.0$, mean: 0.31, average: 0.72.
- Ignition factor: Electrical (EL), Unknown (UNK), Mechanical (ME), Negligence (NE), Chemical (CH), etc. (arson, gas leak (explosion), traffic accidents, and natural factors)
- Ignition material: Unknown (UNK), Electrical (EL), Synthetic Resin (SR), paper and wood (P&W), waste (W), Hazardous Material (HM), etc. (fabrics, food, furniture, gas, signboards, and automobiles).
- Ignition point: Living Space (LS), Facilities and Storage (FS), Function (FN), Structure (STR), Exit (Ex), Process Facility (PF), and Unknown (UNK).

Figure 4 shows the distribution and frequency of the independent variables described above.

The scenario variables to be used for learning are as follows. Regression analysis was conducted on the fire damage ratings included in the input and output information to identify probabilistic correlations between the input and output variables used for learning. For the regression analysis, IBM Statistics (v29.0.1.0) was used.

Among the fire scenario input variables shown in Table 8, except for the season of fire and time of fire, the ignition factor, ignition material, ignition point, human casualties, and burnt area/TFA were found to significantly affect the output variable. It was shown that the season of fire probabilistically had no significant influence compared to other variables because fire evenly occurred regardless of the season.

Table 8. Regression analysis on input variables for output variables (fire scenario).

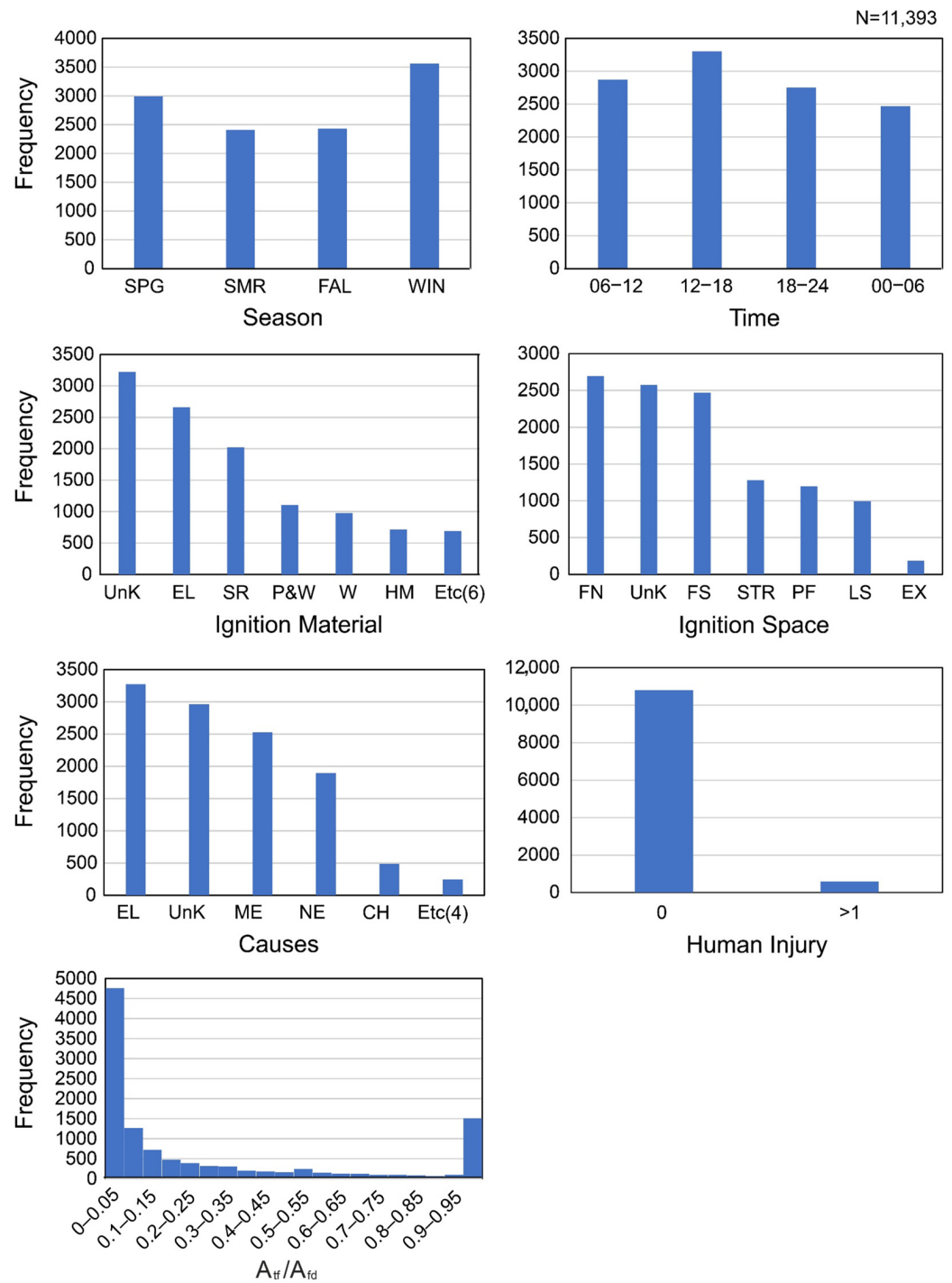|	**Learning Model**	**β**	**SE**	**F**	***p*-Value**
	Season of fire	0.010	0.038	0.076	0.783
	Time of fire	0.039	0.030	1.699	0.193
	Ignition factor	0.190	0.031	37.494	<0.001
Fire scenario	Ignition material	0.107	0.026	17.325	<0.001
	Ignition point	0.208	0.023	82.593	<0.001
	Human casualties	0.096	0.025	14.844	<0.001
	Burnt area/TFA	0.335	0.026	161.88	<0.001

Figure 4. Distribution and frequency of fire scenario information among fire data.

However, ML models were developed using all selected data without probabilistic judgments among the data.

The performance of each ML model was evaluated in further detail using the confusion matrix (Figure 5). The figure below shows the confusion matrix of the training and validation datasets that can compare the actual and predicted values.

It was found that the ANN optimization ML model (classifier) developed earlier well predicted the severe and low ratings compared to the moderate ratings because the prediction levels of the severe and low ratings were nearly identical for both the training and validation models.
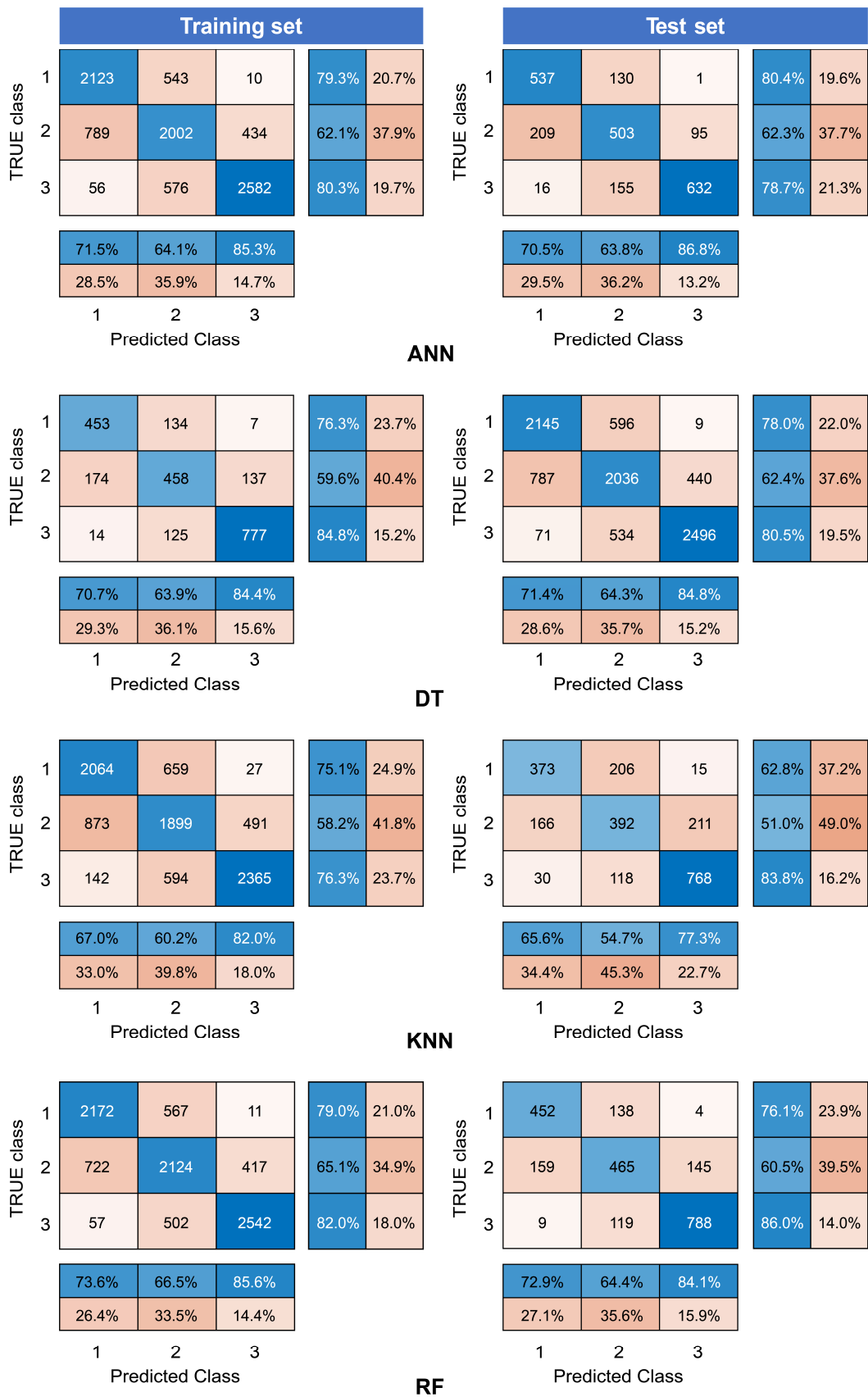
Figure 5. Confusion matrix of ML models (building register information and fire scenario).

For both DT optimization (classifier) and RF optimization ML models, the prediction levels of the severe and low ratings were higher than those of the low ratings. For the validation models, the prediction level of the severe rating was 8–10% higher than that of the low rating.

Overall, the KNN optimization ML model (classifier) exhibited a lower predictive performance than other models. In particular, the predictive performance of the validation model for the moderate and low ratings was lower than that of the training model.

In addition, precision and recall were compared and analyzed to select a classifier learning model based on the values calculated from the confusion matrix.

For the ANN, DT, and RF models, the precision and recall values for the moderate damage rating could not reach the average values, and those for the low and severe damage ratings were slightly higher than the average values.

In the case of the KNN model, the precision and recall values for the low and moderate damage ratings could not reach the average values, and those for the severe damage ratings were slightly higher than the average values.

Figure 5 emphasizes the importance of dividing data into training and test sets. If model training is performed only based on the entire dataset, satisfactory performance cannot be obtained for unknown data (for example, DT, KNN, and RF models yield lower recall for the test set compared to the training set).

4.3. Discussion

Table 9 shows the precision, recall, and F1-score values by rating and the average values based on the validation dataset for the ML models (building register information). F1-score is the harmonic average of precision and recall.

Table 9. Summary of overall ML confusion matrix results (building register information).

	Learning Model	Precision	Recall	F1-Score
ANN	Low	0.430	0.406	0.417
	Moderate	0.371	0.273	0.315
	Severe	0.449	0.595	0.512
	Overall (average)	0.417	0.425	0.421
DT	Low	0.395	0.282	0.329
	Moderate	0.370	0.133	0.196
	Severe	0.452	0.793	0.576
	Overall (average)	0.406	0.403	0.404
KNN	Low	0.387	0.338	0.361
	Moderate	0.346	0.292	0.317
	Severe	0.462	0.568	0.510
	Overall (average)	0.399	0.399	0.399
RF	Low	0.431	0.198	0.271
	Moderate	0.342	0.191	0.245
	Severe	0.446	0.795	0.571
	Overall (average)	0.406	0.395	0.400

Finally, the actual applicability of the ML models was examined by analyzing the confusion matrix results for the validation data.

Overall, the precision, recall, and F1-score values of the severe rating were higher than the average values of each classifier model. This indicates that the prediction success rate for severe fire damage is higher than that for lower damage ratings. In addition, precision and recall were found to be similar.

The ANN classifier model exhibited the highest performance based on the precision, recall, and F1-score averages.

However, considering that the ML models were developed to predict fire damage in advance using simple information, a model with conservative predictions is considered

suitable. Therefore, utilizing the RF model with the highest recall for the severe level can be reasonable.

As the ML models trained only with building register information exhibited precision, recall, and F1-score values of less than 50%, fire scenario data were included in Section 4.3 to develop ML models and examine their performance.

Table 10 shows analyzing the confusion matrix results for the validation data for the ML models (building register information and fire scenario).

Table 10. Summary of overall ML confusion matrix results (building register information and fire scenario).

	Learning Model	Precision	Recall	F1-Score
ANN	Low	0.705	0.804	0.751
	Moderate	0.638	0.623	0.631
	Severe	0.868	0.787	0.826
	Overall (average)	0.737	0.738	0.738
DT	Low	0.707	0.763	0.734
	Moderate	0.639	0.596	0.616
	Severe	0.844	0.848	0.846
	Overall (average)	0.730	0.736	0.733
KNN	Low	0.656	0.628	0.641
	Moderate	0.547	0.510	0.528
	Severe	0.773	0.838	0.804
	Overall (average)	0.659	0.698	0.659
RF	Low	0.729	0.761	0.745
	Moderate	0.644	0.605	0.624
	Severe	0.841	0.860	0.851
	Overall (average)	0.738	0.742	0.740

Overall, the precision, recall, and F1-score values of the severe rating were higher than the average values of each classifier model. This indicates that the prediction success rate for severe fire damage was higher than that for lower damage ratings. In addition, precision and recall were found to be similar.

The RF classifier model exhibited the highest performance based on the average precision, recall, and F1-score values.

RF exhibited the highest precision (73.8%) for the test set, followed by ANN (73.7%) and DT (73%). Additionally, the RF model yielded the highest recall (74.2%), followed by ANN (73.8%) and DT (73.6%). In particular, the recall of the RF model for the severe rating, which is related to one of the important goals of this study, that is, predicting large fires, was 86%. Thus, the RF model exhibited a high overall performance.

It is seen from Figure 6 that the performance of the model varies, depending on the fire scenario. Based on its total accuracy and its fair performance, RF model is suggested as the machine learning model for predicting the fire size. In order to assess the impact of input parameters on the performance of the RF model, additional analysis through a grid search algorithm was conducted to determine the importance of these parameters. This information is shown in Figure 7, where it should be noted that the total sum of all values above the horizontal bars adds up to 100%.

As seen in Figure 7, structure, floor, causes, and total floor area are the critical factors that govern the fire size. It is noted that the burnt area/TFA, fire resistance structure, and season have less influence on the fire size than other parameters.

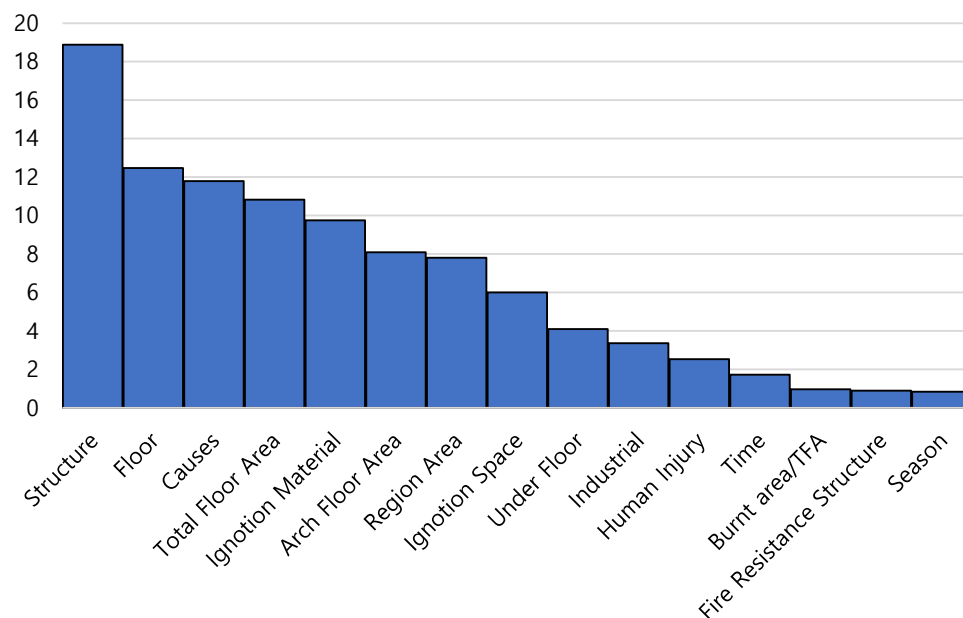**Figure 6.** Comparison of ML models.

Figure 7. Relative importance of input parameters affecting fire size in RF model.

5. Conclusions

As analyzed in Section 2, numerous studies have been conducted to predict the occurrence of fire using various machine learning (ML) methods; however, no methodology exists to predict fire damage ratings only through simple building information. Predicting fire damage using simple data can be effectively used for national and regional disaster management [39]. In this study, the capabilities of ML and artificial intelligence (AI) were explored in identifying the property damage ratings caused by factory fires.

First, a database was constructed by utilizing and preprocessing the fire data provided by a national agency. In the database, 15 input parameters that can predict fire damage ratings based on the insight from past studies were generated and are as follows: facility location information, industry type, structure, fire-resistant structure, number of floors, number of basement floors, total floor area (TFA), building area, burnt area/TFA, season of fire, time of day, ignition factor, ignition material, ignition point, and human casualties. In addition, to increase the prediction and learning success rate, which is the output data, using the 15 input data, the distribution of each rating was analyzed, and the dependent

variable was classified according to the property damage such that a distribution rate of approximately 33% could be obtained.

The entire dataset was divided into training and test sets. The training set was used to set a prediction model, and the model's performance was evaluated through the test set. In this study, four ML models, ANN, KNN, DT, and RF, were evaluated. The performance of the models was evaluated using precision, recall, and F1-score. First, learning (a total of 12,223 data items) was performed using only building register information, and then learning (a total of 11,393 data items) was performed by adding fire scenario information to the building register information to examine the difference.

The performance of the four ML models that performed learning using only building register information was less than 50%; however, the ML performance significantly improved when the four models were trained by adding fire scenario information. Among them, RF exhibited the highest accuracy for the training set, followed by ANN. However, it is difficult to predict fire damage. The proposed RF model showed a recall of 74.2% and a precision of 73.8% in identifying the degree of fire damage for the test set. Notably, it exhibited the highest recall (86%) for the severe rating among the four models. Thus, this learning model can prevent severe property damage by predicting large fires with high probability.

This study demonstrates the capabilities of ML models that predict the degree of property damage in the event of a fire. Open-source data-based classification models can be used in fire centers worldwide to rapidly predict property damage. By analyzing the domestic fire incident data and setting up fire scenarios, a machine learning model using the same approach used in this study can be developed. With this model, registry information of buildings where fires have not yet occurred can be used as prediction data in order to derive property damage size. Fire damage prediction helps establish accident prevention strategies regarding disaster management [40]. It is expected that the results of the proposed prediction model will be utilized for fire prevention activities, such as the management of inspection priorities and inspection periods, while considering the fire risk rating of each building during building fire safety inspections.

The novelty of this study lies in the development of ML models able to predict fire damage size quickly using basic information on buildings. To be specific, the accuracy rate of the RF model, which is around 74%, suggests a great potential of investigating large number of buildings swiftly with high probability. The proposed model also has the flexibility to obtain further insight by accommodating new experimental results. The users can update new experimental results by updating the open-source database and executing the model again. In addition, the proposed classification model may help other researchers plan experimental research. For example, it will be possible to set the dependent variable as the number of casualties or burnt area and predict it. Furthermore, this study demonstrated the functions of ML-based classification models that can be used in disaster management areas other than fire.

This study helps advance the field by demonstrating that it is possible to predict property damage ratings caused by fires using simple building information and ML techniques. It provides insights into developing effective disaster management and prevention strategies by enabling rapid prediction of potential property damage in advance. The proposed classification model can be utilized in various applications related to fire safety inspections and resource allocation for firefighting activities.

The limitation of this study is that the amount of property damage was graded and converted into a classification model to increase the prediction rate for property damage. This can act as an obstacle for putting this research into practical use. Further research and data are needed to predict more specific property damage.

References

1. Lomba-Fernández, L.-F.; Hernantes, H.; Labaka, L. Guide for climate-resilient cities: An urban critical infrastructures approach. *Sustainability* **2019**, *11*, 4727. [CrossRef]
2. Fatih, S.; Ömer, K. Modeling forest fire risk based on GIS-based analytical hierarchy process and statistical analysis in the Mediterranean region. *Ecol. Inform.* **2022**, *68*, 101537.
3. Han, Y.-G.; Baek, J.-S. Recognition of the Risk of Heat Waves as a Disaster and Tasks for Seoul, Seoul Health Foundation. *Seoul Health Air Health Policy Trends* **2022**, *45*, 1–9.
4. Rosyidah, A.; Tambunan, L.; Nurdini, A. Vulnerability analysis of fire evacuation at urban kampong using space syntax method, Penggilingan Jakarta as a case study. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1058*, 012008. [CrossRef]
5. Wang, N.; Xu, Y.; Wang, S. Interpretable boosting tree ensemble method for multisource building fire loss prediction. *Reliab. Eng. Syst. Saf.* **2022**, *225*, 108587. [CrossRef]
6. Su, L.; Wu, X.; Zhang, X.; Huang, X. Smart performance-based design for building fire safety: Prediction of smoke motion via AI. *J. Build. Eng.* **2021**, *43*, 102529. [CrossRef]
7. Wang, Z.; Zhang, T.; Wu, X.; Huang, X. Predicting transient building fire based on external smoke images and deep learning. *J. Build. Eng.* **2022**, *47*, 103823. [CrossRef]
8. Naser, M.Z. AI-based cognitive framework for evaluating response of concrete structures in extreme conditions. *Eng. Appl. Artif. Intell.* **2019**, *81*, 437–449. [CrossRef]
9. Yoon, D.W.; Hwang, H.; Pak, T.Y.; Kim, B.T.; Li, X.; Lee, J. Fire risk prediction using building information and machine learning methods. In *Advances in Information and Communication*; FICC 2022; Lecture Notes in Networks and Systems Volume; Arai, K., Ed.; Springer: Cham, Switzerland, 2022; Volume 438, pp. 22–30. [CrossRef]
10. Marijana, L.; Milos, K.; Meri, C.; Ana, T.-G. Application of artificial neural networks in civil engineering. *Tech. Gaz.* **2014**, *21*, 1353–1359.
11. Wu, X.; Park, Y.; Li, A.; Huang, X.; Xiao, F.; Usmani, A. Smart detection of fire source in tunnel based on the numerical database and artificial intelligence. *Fire Technol.* **2021**, *57*, 657–682. [CrossRef]
12. Lakshmisri, S. Risk analysis model that uses machine learning to predict the likelihood of a fire occurring at A given property. *Int. J. Creat. Res. Thoughts (IJCRT)* **2017**, *5*, 959–962.
13. David, R.; Anna, H.; Dan, E. FireCast: Leveraging deep learning to predict wildfire spread. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019.
14. Wan, X.; Cai, J.; Zhang, B.; Xia, X.; Han, J.; Yan, K. Machine learning method for image recognition-based fire detection system. In Proceedings of the 6th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 4–6 March 2022; IEEE Publications: Piscataway, NJ, USA, 2022.
15. Ye, Z.; Hsu, S.; Wei, H. Real-time prediction of structural fire responses: A finite element-based machine-learning approach. *Autom. Constr.* **2022**, *136*, 104165. [CrossRef]
16. Michigoe, Y. A Study on the Residential Evacuation Safety Assessment Method Based on the Concept of Risk. Ph.D. Thesis, University of Kyoto, Kyoto, Japan, 2012.
17. Nwankpa, C.E.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation functions: Comparison of trends in practice and research for deep learning. In Proceedings of the 2nd International Conference on Computational Sciences and Technology, Jamshoro, Pakistan, 17–19 December 2020; pp. 124–133.

18.   Ntinopoulos, N.; Sakellariou, S.; Christopoulou, O.; Sfougaris, A. Fusion of Remotely-Sensed Fire-Related Indices for Wildfire Prediction through the Contribution of Artificial Intelligence. *Sustainability* **2023**, *15*, 11527. [CrossRef]
19.   Ishola, A.A.; Valles, D. Enhancing Safety and Efficiency in Firefighting Operations via Deep Learning and Temperature Forecasting Modeling in Autonomous Unit. *Sensors* **2023**, *23*, 4628. [CrossRef]
20.   Gabhane, L.R.; Kanidarapu, N. Environmental Risk Assessment Using Neural Network in Liquefied Petroleum Gas Terminal. *Toxics* **2023**, *11*, 348. [CrossRef]
21.   Pang, Y.; Li, Y.; Feng, Z.; Feng, Z.; Zhao, Z.; Chen, S.; Zhang, H. Forest Fire Occurrence Prediction in China Based on Machine Learning Methods. *Remote Sens.* **2022**, *14*, 5546. [CrossRef]
22.   Garcia Leiva, R.; Fernandez Anta, A.; Mancuso, V.; Casari, P. A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *IEEE Access* **2019**, *7*, 99978–99987. [CrossRef]
23.   Hao, Y.; Li, M.; Wang, J.; Li, X.; Chen, J. A High-Resolution Spatial Distribution-Based Integration Machine Learning Algorithm for Urban Fire Risk Assessment: A Case Study in Chengdu, China. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 404. [CrossRef]
24.   Wu, X.; Zhang, G.; Yang, Z.; Tan, S.; Yang, Y.; Pang, Z. Machine Learning for Predicting Forest Fire Occurrence in Changsha: An Innovative Investigation into the Introduction of a Forest Fuel Factor. *Remote Sens.* **2023**, *15*, 4208. [CrossRef]
25.   Sikuzani, Y.U.; Mukenza, M.M.; Malaisse, F.; Kaseya, P.K.; Bogaert, J. The Spatiotemporal Changing Dynamics of Miombo Deforestation and Illegal Human Activities for Forest Fire in Kundelungu National Park, Democratic Republic of the Congo. *Fire* **2023**, *6*, 174. [CrossRef]
26.   Tan, C.; Feng, Z. Mapping Forest Fire Risk Zones Using Machine Learning Algorithms in Hunan Province, China. *Sustainability* **2023**, *15*, 6292. [CrossRef]
27.   Zhang, S.; Cheng, D.; Deng, Z.; Zong, M.; Deng, X. A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognit. Lett.* **2018**, *109*, 44–54. [CrossRef]
28.   Miriam, S.S.; Jastin, P.S.; Pedro, H.A.; Helder, A.; Joao, S. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Comput. Intell. Mag.* **2018**, *13*, 59–76.
29.   Pacheco, A.d.P.; Junior, J.A.d.S.; Ruiz-Armenteros, A.M.; Henriques, R.F.F. Assessment of k-Nearest Neighbor and Random Forest Classifiers for Mapping Forest Fire Areas in Central Portugal Using Landsat-8, Sentinel-2, and Terra Imagery. *Remote Sens.* **2021**, *13*, 1345. [CrossRef]
30.   Heidari, H.; Arabi, M.; Warziniack, T. Effects of Climate Change on Natural-Caused Fire Activity in Western U.S. National Forests. *Atmosphere* **2021**, *12*, 981. [CrossRef]
31.   Angelo, J.J.; Duncan, B.W.; Weishampel, J.F. Using Lidar-Derived Vegetation Profiles to Predict Time since Fire in an Oak Scrub Landscape in East-Central Florida. *Remote Sens.* **2010**, *2*, 514–525. [CrossRef]
32.   Cha, Z.; Yunqian, M. *Ensemble Machine Learning: Methods and Applications*; Springer: Berlin/Heidelberg, Germany, 2012.
33.   Prajwala, T.R. A comparative study on decision tree and random forest using R Tool. *Int. J. Comput. Commun. Eng.* **2015**, *4*, 196–199.
34.   Erenel, Z.; Altınçay, H. Improving the precision-recall trade-off in undersampling-based binary text categorization using unanimity rule. *Neural Comput. Appl.* **2013**, *22*, 83–100. [CrossRef]
35.   Aksoy, E.; Kocer, A.; Yilmaz, İ.; Akçal, A.N.; Akpinar, K. Assessing Fire Risk in Wildland–Urban Interface Regions Using a Machine Learning Method and GIS data: The Example of Istanbul's European Side. *Fire* **2023**, *6*, 408. [CrossRef]
36.   Su, J.; Liu, Z.; Wang, W.; Jiao, K.; Yu, Y.; Li, K.; Lü, Q.; Fletcher, T.L. Evaluation of the Spatial Distribution of Predictors of Fire Regimes in China from 2003 to 2016. *Remote Sens.* **2023**, *15*, 4946. [CrossRef]
37.   Jin, Y.; Zhu, J.; Cui, G.; Yin, Z.; Zhu, W.; Lee, D.K. Characterization of Two Main Forest Cover Loss Transitions in North Korea from 1990 to 2020. *Forests* **2023**, *14*, 1966. [CrossRef]
38.   Qi, Y.; Xue, K.; Wang, W.; Cui, X.; Liang, R. Prediction Model of Borehole Spontaneous Combustion Based on Machine Learning and Its Application. *Fire* **2023**, *6*, 357. [CrossRef]
39.   Sun, W.; Bocchini, P.; Davison, B.D. Applications of artificial intelligence for disaster management. *Nat. Hazards* **2020**, *103*, 2631–2689. [CrossRef]
40.   Lim, D.; Na, W.; Hong, W.; Bae, Y. Development of a fire prediction model at the urban planning stage: Ordinary least squares regression analysis of the area of urban land use and fire damage data in South Korea. *Fire Saf. J.* **2023**, *136*, 103761. [CrossRef]