forests

MDPI

Article

A Forest Fire Prediction Model Based on Meteorological Factors and the Multi-Model Ensemble Method

Seungcheol Choi [1], Minwoo Son [2], Changgyun Kim [3,*] and Byungsik Kim [4,*]

[1] AI for Climate & Disaster Management Center, Kangwon National University, Samcheok 25913, Republic of Korea; tmdak781@kangwon.ac.kr
[2] Department of Urban and Environmental and Disaster Management, Graduate School of Disaster Prevention, Kangwon National University, Samcheok 25913, Republic of Korea; alsdnworks@gmail.com
[3] Department of Artificial Intelligence & Software, Kangwon National University, Samcheok 25913, Republic of Korea
[4] Department of Artificial Intelligence & Software, Graduate School of Disaster Prevention, Kangwon National University, Samcheok 25913, Republic of Korea
* Correspondence: tiockdrbs@kangwon.ac.kr (C.K.); hydrokbs@kangwon.ac.kr (B.K.)

Abstract: More than half of South Korea's land area is covered by forests, which significantly increases the potential for extensive damage in the event of a forest fire. The majority of forest fires in South Korea are caused by humans. Over the past decade, more than half of these types of fires occurred during the spring season. Although human activities are the primary cause of forest fires, the fact that they are concentrated in the spring underscores the strong association between forest fires and meteorological factors. When meteorological conditions favor the occurrence of forest fires, certain triggering factors can lead to their ignition more easily. The purpose of this study is to analyze the meteorological factors influencing forest fires and to develop a machine learning-based prediction model for forest fire occurrence, focusing on meteorological data. The study focuses on four regions within Gangwon province in South Korea, which have experienced substantial damage from forest fires. To construct the model, historical meteorological data were collected, surrogate variables were calculated, and a variable selection process was applied to identify relevant meteorological factors. Five machine learning models were then used to predict forest fire occurrence and ensemble techniques were employed to enhance the model's performance. The performance of the developed forest fire prediction model was evaluated using evaluation metrics. The results indicate that the ensemble model outperformed the individual models, with a higher F1-score and a notable reduction in false positives compared to the individual models. This suggests that the model developed in this study, when combined with meteorological forecast data, can potentially predict forest fire occurrence and provide insights into the expected severity of fires. This information could support decision-making for forest fire management, aiding in the development of more effective fire response plans.

Keywords: forest fire; meteorological factor; multi-model ensemble; machine learning

Citation: Choi, S.; Son, M.; Kim, C.; Kim, B. A Forest Fire Prediction Model Based on Meteorological Factors and the Multi-Model Ensemble Method. *Forests* **2024**, *15*, 1981. https://doi.org/10.3390/f15111981

Academic Editors: Rafael De Ávila Rodrigues and Rafael Coll Delgado

Received: 22 September 2024
Revised: 2 November 2024
Accepted: 4 November 2024
Published: 9 November 2024

1. Introduction

More than half (62.6%) of Korea's land area is covered by forests [1]. When widely distributed forests extensively cover a significant portion of a country, such as Korea, the country is highly susceptible to fire hazards. This susceptibility to fire hazards can result in substantial damage to forests due to forest fires [2]. In addition, in Korea, the percentage of mountainous terrain is higher than that of flat land, and the forests are also densely distributed, increasing the likelihood of large-scale forest fires [3]. Regions affected by forest fires suffer a depletion in forest resources, rendering them more vulnerable to disasters, such as landslides. Such depletion may even lead to a decrease in local biodiversity. The resultant ecological and scenic damage induce secondary damage, such as lower tourism

demand [4]. In Korea, forest fires predominantly occur in March–May, with spring forest fires accounting for 85% of the total affected area [5]. As such, South Korea Forest Service (KFS) has designated the period between February 1 and May 15 as the forest fire watch period, implementing measures, including access control to mountains and setting up forest fire watchhouses at various locations. According to the KFS's forest fire status, 523 forest fires occurred in the 2000s, which increased to 575 in the 2020s (2020–2022). The total area affected by forest fires in the 2000s was 3726 ha, which tripled in the 2020s, reaching 9494 ha. The increase in the area affected by forest fires can be attributed to more frequent dry days and stronger winds induced by climate change, leading to larger forest fires than in previous years [6]. Moreover, 84.3% of forest fires in South Korea as of 2018 were caused by human activities [6]. Of these, 53.5% were the result of the direct handling of fire in mountainous areas, such as part of rituals and smoking. In contrast, forest fires caused by weather factors are mainly caused by humidity and wind. Strong winds can carry embers over long distances and the dry climate accelerates the spread of forest fires. In South Korea, 10.5% of forest fires were caused by wind-blown embers originating from outside the forest and 18.1% were caused by embers from agricultural activities. Combined, these two factors account for 28.6% of forest fires and are major contributors to the size and severity of forest fires. Among the weather factors, lightning as a direct cause accounted for only 61 cases or 0.5% of all forest fires from 1990 to 2018. Given this low frequency of occurrence, lightning is not considered a major cause of forest fires in South Korea [7].

Forest fires tend to surge in spring and autumn due to rising temperatures and prolonged dry periods [8]. The purpose of this study is to develop an advanced prediction model that accurately forecasts the monthly occurrence of forest fires, by incorporating region-specific seasonal meteorological conditions. As meteorological factors are critical determinants of forest fire occurrence, this study focuses on enhancing the model's prediction accuracy through the application of machine learning techniques combined with these key variables. To mitigate the potential shortcomings of single-model approaches, which may lead to prediction failures in certain cases, a multi-model ensemble framework is proposed. By employing ensemble methods to integrate multiple models, this research seeks to substantially improve the model's prediction performance and establish a more reliable and robust forecasting model.

The organization of this paper is as follows: Section 2 introduces various studies conducted on forest fire prediction and discusses their implications. Section 3 outlines the methodology in this study, including the machine learning techniques employed. Section 4 details the data processing steps taken to prepare the training dataset for use in the machine learning model. In Section 5, we present a comparative analysis of the prediction results from the MME-based model proposed in this study and the results from a single model, using graphs and evaluation metrics. Section 6 discusses the findings presented in Section 5 and, finally, Section 7 provides a summary of the study and suggests possible directions for future research.

2. Literature Review

Numerous studies have been conducted in order to find ways to mitigate the considerable damage caused by forest fires. Classical methods for predicting forest fire risk include the development of formula-based forest fire risk indices. The oldest record found in regard to this area of study is the Munger Index from 1916 [9] and it is assumed that similar attempts have been made steadily over the years, in various parts of the world. Many scholars have emphasized that weather is a critical factor in regard to forest fire occurrence [10]. For this reason, a variety of forest fire risk indices, up to the present day, use climatic or weather conditions to predict fire occurrence, often incorporating topographical elements. Fuel indices typically consider the fuel layer of the land, heavily factoring in the duff moisture and land cover, with fuel dryness regarded as the greatest threat [11]. Second, indices reflecting the risk of fire occurrence and the conditions that could contribute to fire severity are incorporated. Triggers for forest fires can be anthropogenic, but they can also occur

without any human involvement, such as through lightning strikes [12–14]. Both socio-environmental and natural factors, such as weather, can be considered. However, from a community perspective, human activity tends to follow patterns that can be quantified using specific coefficients. These indices can exhibit different characteristics, depending on the cultural and climatic conditions of each region. Consequently, each country has developed its own unique forest fire risk index. While the core elements of these indices may be similar, the specific coefficients often reflect the characteristics of each nation [15].

Historically, these unique coefficients were based on empirical formulas derived from numerical analysis and modeling. Representative techniques include the KBDI (Keetch–Byram Drought Index) [16], the FWI System (Canadian Forest Fire Weather Index System) [17], the NFDRS [18], and the Nesterov ignition index [19]. In addition to using simple indices to calculate the forest fire risk, studies have also been conducted to predict the number of forest fire occurrences. However, due to the high level of nonlinearity involved, research on predicting the number of forest fires has been relatively underdeveloped compared to studies on forest fire risk indices. Recently, studies on predicting the number of forest fires have been actively conducted. These studies aim to use artificial intelligence to establish hidden weights for the socio-environmental factors related to specific regions and seasons and to predict forest fire occurrence or probabilities through the use of forest fire risk indices. Ref. [20] predicted forest fire areas using forest fire data from Portugal's Montesinho Natural Park and five components from the Fire Weather Index (FWI) System and weather factors. The optimal performance was achieved using a support vector machine (SVM) and weather factors (temperature, rain, relative humidity, and wind velocity). In 2018, ref. [21] also predicted forest fire areas in the same target area using the random forest (RF) method, with an RMSE of 8.37 in regard to the prediction performance, while the XGBoost algorithm exhibited the highest prediction accuracy (72.3%) for the large forest fire classification. Ref. [22] proposed a sparse autoencoder-based deep neural network (DNN) method for large forest fire prediction, coupled with a data balancing method to address the imbalance issues, reporting the lowest RMSE (0.95–19.3) among the ANNs (artificial neural networks), namely SVM and RF methods. Ref. [23] utilized fuzzy inductive reasoning (FIR) to identify the most relevant features for predicting the forest fire area in Montesinho Natural Park, finding a strong causal relationship between weather factors and the Fine Fuel Moisture Code (FFMC). In 2019, ref. [24] used forest fire area, duration, and weather factors to predict the size of forest fires in Alberta, Canada. As for the weather factors, a variance inflation factor (VIF) was applied to remove the minimum temperature, mean temperature, and total precipitation variables and compared the results of the back-propagation neural network (BPNN), recurrent neural network (RNN), and long short-term memory (LSTM). In the study, the LSTM network showed the highest prediction accuracy of 90.9%. Ref. [25] developed a prediction model for forest fire frequency, utilizing a deep belief network (DBN), with an NSE and an RMSE of 0.87 and 0.07, respectively. The study used average weather conditions and forest fire frequency for the Korean Peninsula. In addition to numerical data, such as meteorological factors, research has also explored the prediction of forest fires using non-numerical data. Ref. [26] used various technologies for forest fire prevention that have been introduced to enable effective prevention and warning measures. Ref. [27] developed the FlameTransNet, which leverages the Transformer module and CBAM (Convolutional Block Attention Module) for early warnings and rapid responses to forest fires. FlameTransNet outperformed the UNet model. Ref. [28] proposed a system that integrates audio and image data for forest fire detection. This system employed datasets that included audio and image recordings from both fire and non-fire scenarios. Ref. [29] introduced a deep learning-based fire detection system, utilizing the YOLO-v8 algorithm and achieving a high accuracy rate of 97.1%. Ref. [30] investigated the Dr-TOBID system, which employs drones and deep learning to detect smoke and flames across varying altitudes, at all times of day. Finally, ref. [31] presented the MD-CNN (modified deep convolutional neural network) method, based on transfer learning and a feature fusion algorithm, for rapid forest fire detection, which demonstrated a high

level of accuracy of 95.8% and a recall rate of 95.4%. There are systems in place that utilize both numerical and non-numerical data to prevent forest fires. According to [32], Europe currently operates the European Forest Fire Information System (EFFIS), which utilizes a combination of meteorological and satellite data. The EFFIS is a system developed to support forest fire prevention and responses in Europe, and it predicts forest fire risk based on the FWI System and European meteorological data and detects large-scale forest fires in real-time through the use of NASA satellite data (MODIS).

The existing studies on forest fire prediction allow us to make several conclusions. Weather factors have been pivotal in forest fire prediction and have been given great importance in the literature, with the Canadian Forest Fire Index (FWI) System being widely utilized. Techniques, such as the data balancing method, have been applied to solve the data imbalance in forest fire occurrence data. VIF analysis has been employed to address data imbalances and multicollinearity issues among independent variables. A lot of research has been conducted on machine learning-based forest fire prediction, but such research is mainly related to predicting the areas affected by forest fires [33–40]. Predicting the areas likely to be affected by forest fires can help minimize the damage they cause. However, such research is far from being able to prevent forest fires. Therefore, there is a need to provide local government forest fire managers with intuitive information on the likely frequency of forest fires in the future. This will allow them to take early preventative measures to protect forest resources from forest fires.

3. Methodology

3.1. The Study Flowchart

The development procedure for the forest fire prediction model proposed in this study is shown in Figure 1, and the model was developed in three steps. The model developed in this study is a model that predicts the occurrence of forest fires and non-forest fires on a daily basis, and then adds up the prediction results to predict the number of forest fires per month. First, during the data collection and processing stage, the processing of the collected hourly data into daily data takes place, the proxy variables are calculated using meteorological factors, and the data processing using box plots is performed. Second, during the selection of independent variables stage, variable selection techniques are applied to select the variables, multicollinearity among the selected variables is checked, and the SMOTE algorithm is applied to prepare the training data. Finally, during the multi-model ensemble method development stage, the input data preprocessing process and the machine learning model development process are performed. The concept of the multi-model ensemble method is shown in Figure 2. Five models were selected, based on the preliminary validation results on their predictive performance. During this process, different models, such as the Naive Bayes and CatBoost algorithms, were rigorously evaluated for their consistency and accuracy. The models that demonstrated reliable and robust performance were included in the final model selection. After training five different models, the prediction results from each model are compared. If four or more of the five models predict a forest fire, it is finally determined that a forest fire has occurred. To reflect the seasons, a 'Season' column is added, which is divided into spring, from March to May, summer, from June to August, fall, from September to November, and winter, from December to February, and one-hot encoding is applied, as shown in Figure 3.

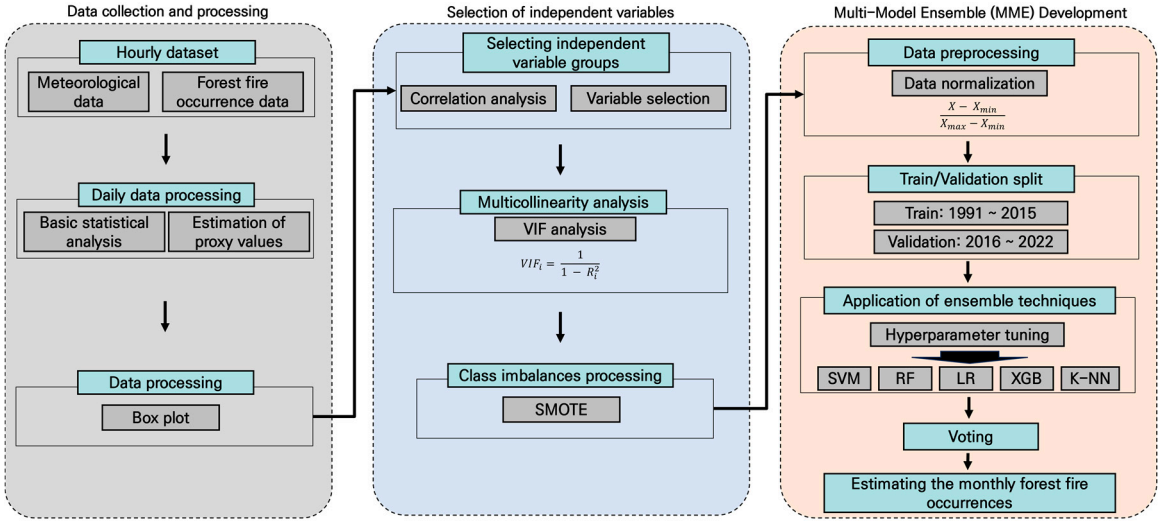


Figure 1. Research flowchart.

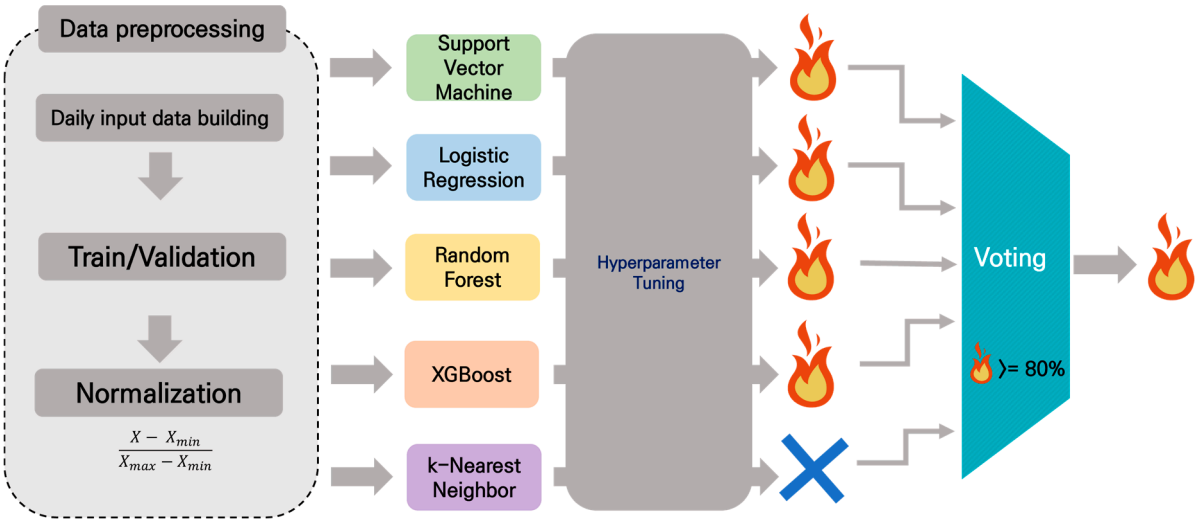


Figure 2. The concept of the multi-model ensemble method. If at least 4 out of 5 models predict a forest fire, this model finally predicts that a forest fire has occurred.

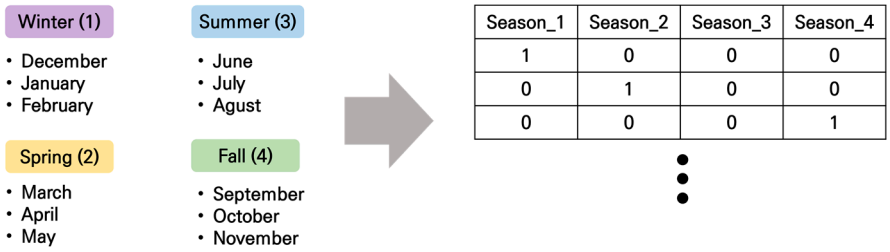


Figure 3. Applying one-hot encoding to the "Season" column.

3.2. Extreme Gradient Boosting (XGB)

The Extreme Gradient Boosting (XGB) algorithm, proposed by [41] in 2016, is a notable advancement in machine learning, particularly for regression and classification problems. For a theoretical understanding and mathematical formulation of XGB, the works of [41–43] are valuable. The algorithm employs K-additive functions to predict outcomes for n training instances, each with m features. This approach, encapsulated in Equation (1),

leverages the power of ensemble learning, by combining multiple models to significantly improve the predictive performance of the model.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F} \tag{1}$$

where $\mathcal{F}$ is the function space to which f_k belongs, which is defined as $\mathcal{F} = \left\{ f(\mathbf{x}) = \omega_{q(\mathbf{x})} \right\}$ $\left(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T\right)$; $q(\mathbf{x})$ is a tree structure of which the decision rule maps the input data $\mathbf{x}$ to the corresponding leaf index; $\omega_{q(\mathbf{x})}$ is the weight of the leaf mapped according to which the input data $\mathbf{x}$ are mapped using decision rule $\mathbf{x}$, with the decision rule $q(\mathbf{x})$ mapped to the weights of the leaves using decision rule $q(\mathbf{x})$; T is the number of leaves on the tree, f_k is the k-th classification and regression tree (CART) model, with a tree structure; $\hat{y}_i$ is the predicted value of the i-th sample; K is the total number of trees; and $\mathbf{x}_i$ is the i-th input data. The objective function of XGB is constructed as shown in Equation (2).

$$\begin{gathered} L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ \text{where, } \Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2 \end{gathered} \tag{2}$$

where l is a differentiable convex loss function that calculates the difference between y_i and $\hat{y}_i$; γ is the complexity of each leaf; λ is the parameter that adjusts the size of the penalty; ω_j is the weight of the j-th leaf; and Ω is a normalization term that smooths out the final learned weights to avoid overfitting.

XGBoost offers a variety of hyperparameters that can be adjusted, such as n_estimators, which controls the number of trees (K) to be generated; learning_rate, which controls how much the model's weights are reduced at each learning step; max_depth, which determines the maximum depth of the trees; and gamma, which regulates the minimum γ in terms of the splits. In this study, hyperparameter tuning was conducted to avoid overfitting and to develop an optimal model.

3.3. Random Forest (RF)

Random forest (RF) is a method used to create and train multiple decision tree models, which then combines the predictions from each tree to produce a single, more accurate prediction [44]. RF applies bootstrap sampling to the training data, to train a decision tree model based on randomized samples. Generating a single prediction result has the advantage of preventing overfitting by reducing the dependence between the trees, because the average value of the predictions of each tree is used for the regression problem and the most selected class through majority voting is used for the classification problem [45]. In addition, RF offers high prediction accuracy and tolerance to outliers and noise, which results in good forest fire prediction performance [46]. When the input data are x, an RF model's training and prediction process can be described as detailed below.

Bootstrap samples of size S are extracted from the training data of size N for B trees, and algorithms (i) through to (iii) below are repeated using the sample data to develop the random forest tree n_{min} (b = 1, …, B), until the minimum terminal node size n_{min} is obtained.

(i) Randomly select p variables from a total of M variables;
(ii) Select the optimal variable/division point among the p variables;
(iii) Split the node into two child nodes.

The above process results in an ensemble tree $\{T_b\}_{b=1}^{b=B}$. As for the classification problem, if the class prediction result of the b-th tree is $\hat{C}_b(x)$, (x) is obtained as a result of the majority voting in terms of the B trees $\left\{\hat{C}_b(x)\right\}_1^B$ [47].

3.4. Logistic Regression (LR)

Logistic Regression (LR) is a statistical method for modeling the probability of a binary outcome, based on one or more predictor variables. It is similar to a general regression model in that a linear combination of independent variables explains the dependent variable. It differs from linear regression by predicting a probability that ranges between 0 and 1, thus making it suitable for binary classification tasks. In addition, in a general regression model, the dependent variable ranges from negative infinity to positive infinity, whereas in an LR model, the dependent variable is expressed as an S-shaped function, with a range from 0 to 1 [48].

In this study, the first-level forest fire occurrence forecast results fall within a binary classification, with two classes: forest fire and non-forest fire. In the LR model, for the binary classification (y= 1, 0), if P_i represents the probability of y= 1 for the i-th sample, P_i is calculated as shown in Equation (4).

$$logit(P_i) = log\frac{P_i}{1-P_i} = z_i \tag{3}$$

$$P_i = \frac{1}{1+exp(-z_i)} \tag{4}$$

$$where, z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where $logit(P_i)$ is the logit transformation; exp is a natural constant; k is the number of independent variables; β_0 is bias; β_j is the regression coefficient of the j-th independent variable; and x_{ij} is the j-th independent variable in the i-th sample.

3.5. The k-Nearest Neighbor (k-NN)

The k-NN technique is one of the most popular supervised learning-based machine learning techniques for classification problems [49] and has been applied in various forest fire research fields [49–52]. Given an input x, a k-NN finds the k-closest data points to form a neighborhood with x and classifies x through majority voting [53]. The k-NN can find the best model by adjusting the hyperparameters, such as k, and the distance calculation method, which is used to find the nearest neighbors to x. The k-NN technique can be tuned to find the best model. The classification process of the k-NN technique is shown in Figure 4.

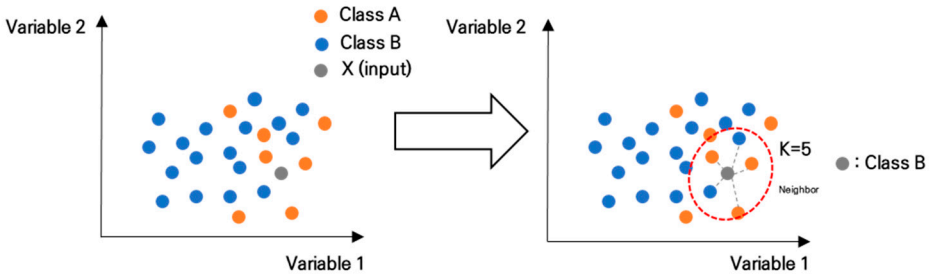


Figure 4. The classification process of the k-Nearest Neighbor (k-NN) technique.

3.6. Support Vector Machine (SVM)

Support vector machine (SVM) is a statistical machine learning method proposed by [54]. Fundamentally, SVM is designed for binary classification tasks, aiming to identify an optimal hyperplane in the feature space that maximizes the margin between the two classes, while allowing for a certain amount of learning error. In this case, the optimal hyperplane is defined by some of the data points in the training data, called support vectors [54]. The theoretical understanding and mathematical formulations of a SVM are elaborated in works such as [55–58]. If the training data were represented by N samples,

n characteristics, and two classes labeled as -1 and 1, respectively, the calculation would be as follows:

$$\{(\mathbf{x}_1, y_1), ...(\mathbf{x}_i, y_i), ..., (\mathbf{x}_N, y_N)\} \text{ where, } \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in} \end{bmatrix}, \ y_i \in \{-1, 1\}$$

When the classes are linearly separable, the optimal hypersurface that separates the two classes can be shown as Equation (5).

$$\mathbf{w}_0^{\mathrm{T}} \cdot \mathbf{x} + b_0 = 0 \tag{5}$$

The optimal hyperplane has the maximum margin, and $\mathbf{w}_0^{\mathrm{T}}$ and b_0 are estimated, so that the data points closest to the hyperplane satisfy the two constraints in Equation (10). The data points used become the support vectors. Equation (6) can be expressed as a single expression, as shown in Equation (7).

$$\begin{cases} \boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \geq 1, & if\ y_i = 1 \\ \boldsymbol{w}^T \cdot \boldsymbol{x}_i + b \leq -1, & if\ y_i = -1 \end{cases} \tag{6}$$

$$y_i\left(\mathbf{w}^{\mathrm{T}} \cdot \mathbf{x}_i + b\right) - 1 \geq 0, i = 1, \ \ldots, \mathrm{N} \tag{7}$$

In this case, the margin is the distance between the two boundaries in Equation (6), expressed as $\frac{2}{\|\mathbf{w}\|}$. This means that the maximum margin is the distance at which $\frac{2}{\|\mathbf{w}\|}$ is maximized, which can again be expressed as Equation (8).

$$min \frac{1}{2} \|\mathbf{w}\|^2 \tag{8}$$

In this study, hyperparameter tuning was performed using a grid search, by adjusting certain parameters, such as C, which controls the regularization strength, and kernel functions, such as the Radial Basis Function (RBF), Gaussian kernel, and polynomial kernel, along with γ, which regulates the width of the kernel functions. The optimal combination of parameters was selected to develop the SVM model with the best performance.

3.7. Performance Evaluation Metrics

The forest fire prediction model proposed in this study predicts the number of forest fires per month by predicting the occurrence of forest fires and non-forest fires on a daily basis and then adding up the prediction results for each month. Therefore, standard evaluation metrics commonly utilized in classification problems were employed to assess the model's performance in regard to predicting the daily occurrence of forest fires. Specifically, the accuracy and F1-score were utilized as evaluation metrics, represented by Equations (9) and (10).

$$Accuracy \ = \ \frac{TP + TN}{TP \ + TN + FP + FN} \tag{9}$$

$$F1 - score \ = \ 2 \ * \ \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{10}$$

Here, TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively. The accuracy measures the proportion of correctly predicted outcomes among all the observations, offering an overall assessment of the predictive performance of the model. On the other hand, the F1-score, derived from the harmonic mean of the precision and recall [59], provides a balanced evaluation metric suitable for imbalanced data.

4. Data Preprocessing

4.1. Target Area

Figure 5 shows the target area of this study. We selected Gangneung, Samcheok, Chuncheon, and Hongcheon, all of which are located in Gangwon province, South Korea. Gangwon-do encompasses 21% of Korea's total forested area, with forests covering a vast 81% of the province. Consequently, it has endured forest fire damage in more regions than any of the other 16 provinces in Korea [60,61]. The Taebaek Mountain Range runs through Gangwon province, dividing the region into Yeongseo to the west and Yeongdong to the east. The term "Yeongseo" refers to the western (西, seo) side of the range (嶺, Yeong) and "Yeongdong" refers to the eastern (東, dong) side. As the entire province of Gangwon is located within the Taebaek Mountain Range, this geographic feature serves as a standard reference in regard to the climate classification. The climate of the Yeongseo region, which includes areas, such as Hongcheon and Chuncheon, is classified as a humid continental climate with dry winters (Dwa), according to the Köppen climate classification. Chuncheon falls within the Dwa category, while Hongcheon is categorized as both Dwa and Dwb. In contrast, the Yeongdong region, which includes Gangneung and Samcheok, features a mix of climates, including some Cfb zones. The coastal areas are classified as Cfa, while the mountainous regions fall within Dfb. Gangneung also contains small pockets of Dfa, while both Gangneung and Samcheok follow the general pattern of Cfa for coastal areas and Dfb for mountainous regions [62]. The forest composition in Gangwon province is also distinguished by the division between the Yeongseo and Yeongdong regions. In the Yeongdong region, coniferous forests are primarily distributed along the eastern coastal areas, extending inland toward the forested areas near Gangneung. Samcheok, on the other hand, is dominated by extensive broadleaf forests in the western mountainous areas. In the western part of the province, the forests of Chuncheon are predominantly small broadleaf forests, while Hongcheon is characterized by a predominance of mixed forests (Figure 6) [60].

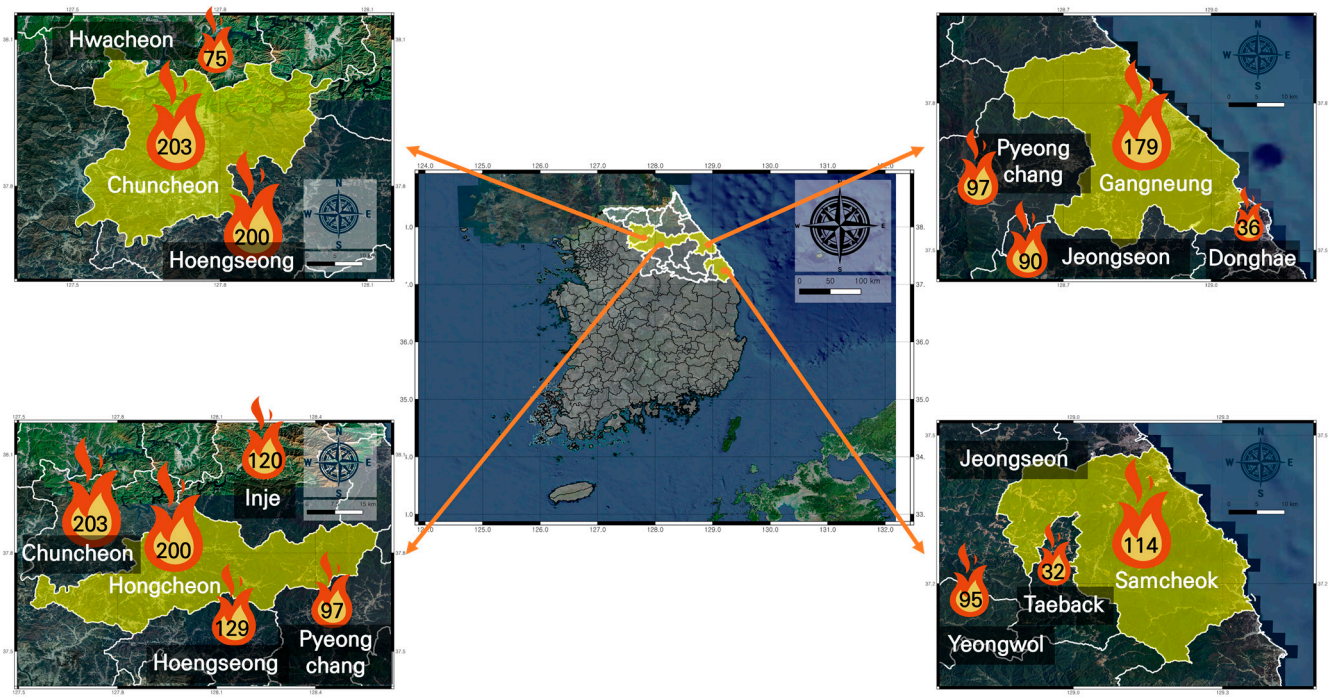


Figure 5. The study's target area. A visualization of the location of Gangwon province in South Korea, the location of the study subject within Gangwon province, and the number of forest fires.

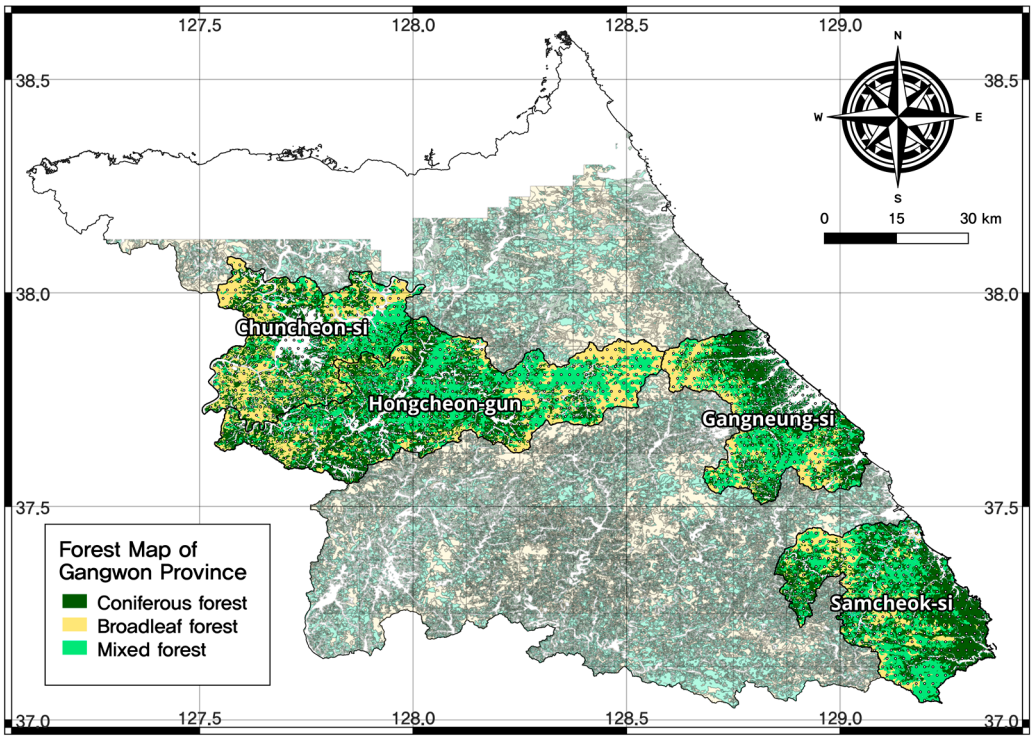


Figure 6. The forest composition [60] in Chuncheon, Gangneung, Hongcheon, and Samcheok is presented, starting from the second quadrant and proceeding clockwise. Dark green represents coniferous forests, light green indicates mixed forests, and yellow represents deciduous forests.

In this study, to account for the distinct characteristics of the Yeongseo and Yeongdong regions mentioned above, two areas with a high frequency of forest fires were selected from each region. To analyze the occurrence of forest fires in the target area and the surrounding areas, we collected forest fire occurrence data from 1991 to 2022, provided by the Korea Forest Service. As can be seen in Figure 5, the analysis shows that all of the target regions had a high number of forest fires compared to neighboring regions. Figure 7 shows the number of forest fires in the target area by year, from 1991 to 2022.

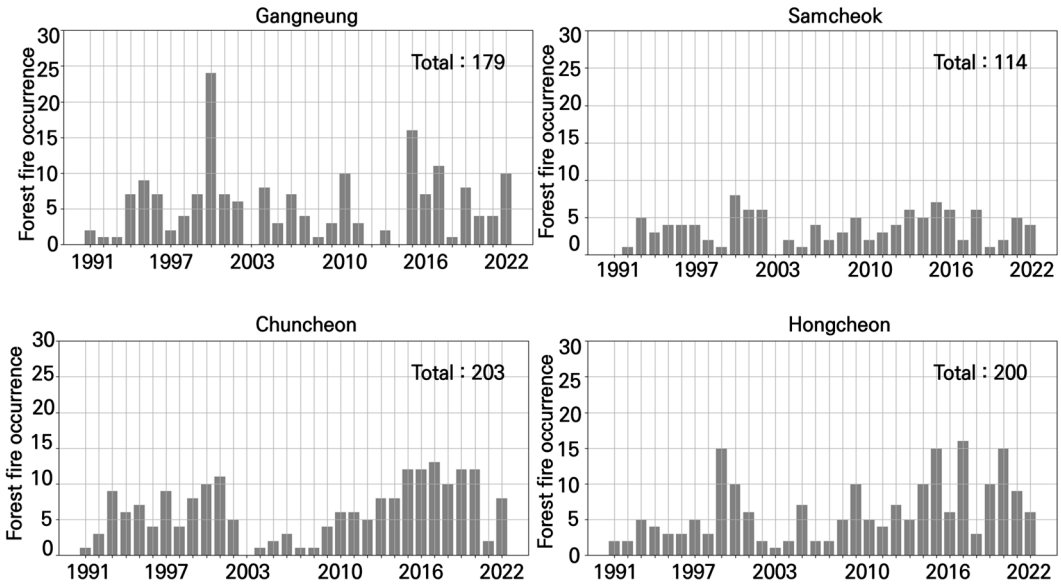


Figure 7. Frequency of forest fires by year in the target areas.

Table 1 presents the data utilized in this study. The collected data consist of hourly observations, which were processed into daily data through the data processing procedure illustrated in Figure 8. During this process, the daily maximum, minimum, and average values were calculated for each variable, along with the derivation in terms of the proxy variables. Additionally, moving averages (3-day, 7-day moving averages) were applied to account for the influence of past meteorological conditions on the present values.

Table 1. Description of the data utilized in the study.

Data	Name	Source	Unit	Period (Year)	Abbreviation
Forest fire data	Forest fire occurrence	Korea Forest Service (KFS)	-	1991~2022	-
Meteorological data	Wind speed	Korea Meteorological Administration (KMA)	m/s		WS
	Temperature		°C		TA
	Relative humidity		%		HM
	Dew point temperature		°C		TD
	Precipitation		mm		PCP
Proxy variables	Fine fuel moisture code	-	-		FFMC
	Effective humidity		%		EFHM
	No precipitation days		day		N_PCP_days

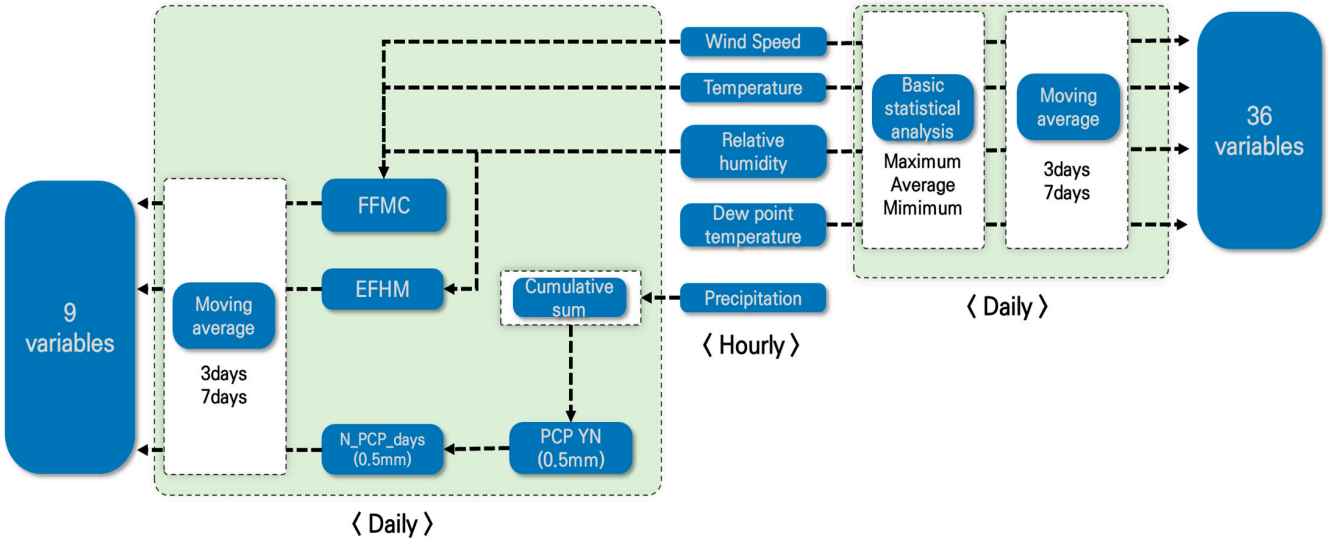


Figure 8. Raw data processing procedure. In order to take into account not only the influence of the meteorological factors on the day, but also the influence of past meteorological factors, the variables were calculated by applying a moving average.

4.2. Proxy Variables

In this study, three proxy variables based on the meteorological factors were calculated. First, the effective humidity (*He*) represents dryness by allocating a weight to the relative humidity from several days before the event, according to the elapsed time. Similar to relative humidity, it ranges from 0 to 100 and is calculated by allocating a weight to the relative humidity from four days before the event, according to the elapsed time, as shown in Equation (11), where r is the coefficient. In this study, 0.7, which is the current

coefficient used by the Korea Meteorological Administration (KMA), was utilized [63–65]. ${H_t}^{0d}$ is the average humidity on the day and ${H_t}^{xd}$ is the relative humidity x days ago.

$$He = (1 - r) \times \left[{H_t}^{0d} + r{H_t}^{1d} + r^2{H_t}^{2d} + r^3{H_t}^{3d} + r^4{H_t}^{4d} \right] \quad (11)$$

Second, the FFMC is one of the main factors in the FWI System, a meteorological index used by the Canadian Forest Fire Danger Rating System (CFFDRS), proposed by [66], which indexes the moisture content of fine fuel on the forest floor using four meteorological factors: air temperature, relative humidity, wind velocity, and daily precipitation [4]. The FFMC ranges from 0 to 99, with a higher number indicating a lower moisture content and a higher forest fire risk.

Finally, the no precipitation days factor represents the number of consecutive days without rain. In this case, if the precipitation was 0.5 mm or less, it was considered as a no precipitation day. A higher number of days without rain lowers the leaf moisture content, fueling forest fires and prolonging dryness in the atmosphere, creating favorable conditions for forest fires. Therefore, we utilized this variable in our study.

4.3. Data Processing by Comparing the Distribution of Forest Fires

Box plots allow a visual comparison of the distribution of data between forest fire and non-forest fire days, and they also identify outliers that fall outside this boundary. Identifying and handling outliers is a crucial step in the data preprocessing stage, as outliers can significantly influence the results of the analysis and can lead to inaccurate conclusions [67]. The boundaries established in this study are defined below. Where Q_1 is the first quartile, Q_3 is the third quartile, IQR is the interquartile range, which is the difference between the third and first quartiles, F_{in} is the inner boundary, and F_{out} is the outer boundary.

$$IQR = Q_3 - Q_1 \quad (12)$$

$$F_{in} = Q_1 - 1.5 * IQR \quad (13)$$

$$F_{out} = Q_3 + 1.5 * IQR \quad (14)$$

The distribution of the data between forest fire and non-forest fire days for the key variables was visualized as a box plot, and the results are shown in Figure 9. The red dots in each plot identify the outliers that fell outside the established boundaries. In Table 2, a class imbalance problem is apparent. In order to develop a machine learning model with good performance in the presence of a class imbalance problem, it was necessary to make the proportion of the two classes similar, as well as to create a clear difference between the data distribution in terms of the forest fire and non-forest fire days. In addition, extreme meteorological conditions (a high FFMC, low effective humidity, high wind speed, etc.) that are prone to causing forest fires had to be preserved. Therefore, in this study, the data processing was conducted considering the aforementioned points. First, the quantiles and fences were used to handle the outliers and to make clear the differences in the distribution of the data between the forest fire and non-forest fire days. In addition, the data points that fell outside the boundaries, but corresponded to extreme meteorological conditions, were identified and preserved as extreme values rather than outliers. Figure 10 shows the before and after data processing results for Gangneung, one of the target areas, as an example. It can be seen that a clear difference exists in regard to the distribution of the data between the forest fire and non-forest fire days. Table 2 presents the results of the data processing by target area, and it can be seen that the data from the forest fire days are preserved as much as possible and the data from the non-forest fire days are removed.

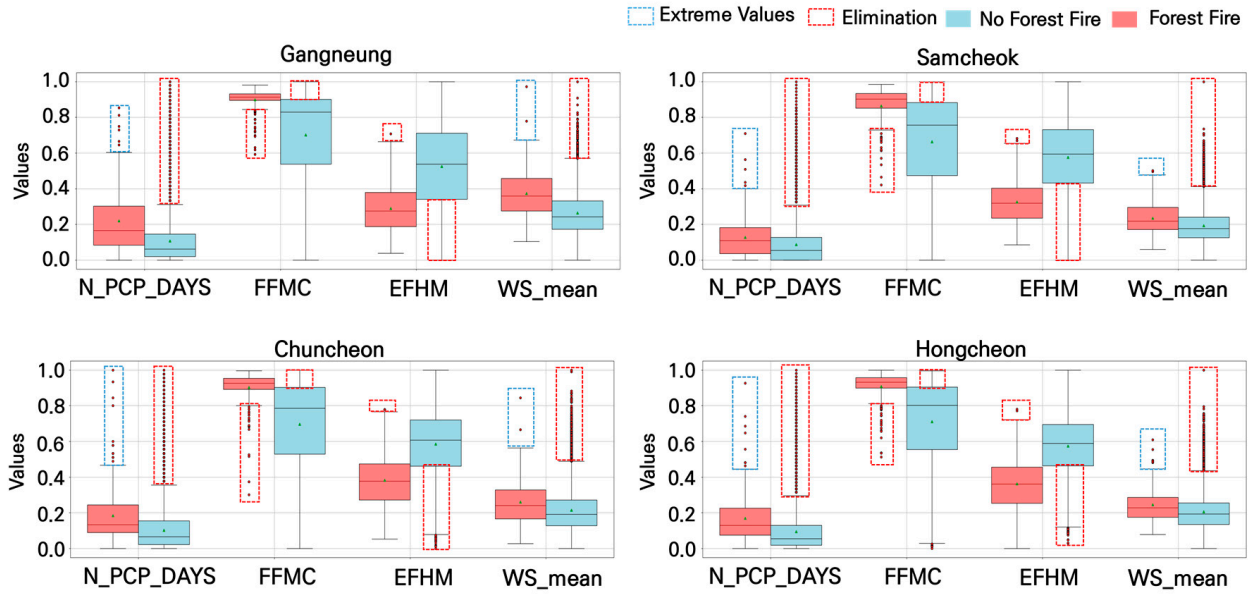


Figure 9. Outliers and extreme values for each variable.

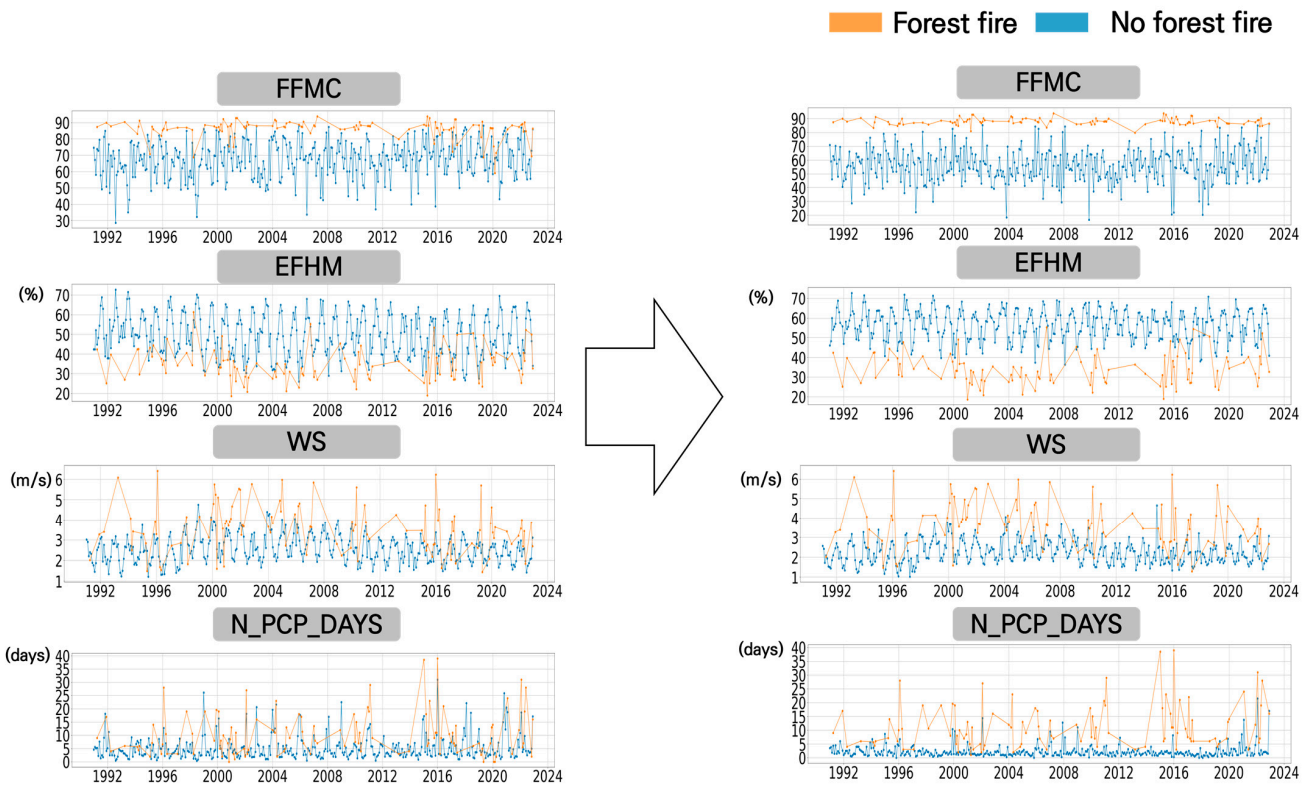


Figure 10. The results of the data processing as a monthly average (example: Gangneung).

Table 2. The results of the data processing by target area. It can be seen that we have deleted Non-Forest Fire, while preserving Forest Fire, as much as possible. However, as Table 2 shows, there is still a class imbalance problem.

Region	Type	Forest Fire	Non-Forest Fire	Total
Gangneung	Before processing	179	11,509	11,688
	After processing	163	6686	6849
Samcheok	Before processing	114	11,574	11,688
	After processing	99	6641	6740
Chuncheon	Before processing	203	11,485	11,688
	After processing	186	5970	6156
Hongcheon	Before processing	200	11,488	11,688
	After processing	171	6431	6602

4.4. Selection of the Independent Variables Through Variable Selection

During model training, the inclusion of variables irrelevant to or sharing overlapping features with dependent variables can compromise the model's performance. Consequently, eliminating irrelevant features can enhance the model's accuracy and reduce the model training time by reducing the data dimensions [68]. In this study, the variables were selected in two stages. In the first step, the forward selection method was applied to select the variables that were highly correlated with the dependent variable from the total variables. In the second step, the selected variables were grouped together with the variables that had the same characteristics to create variable groups. Then, the correlation coefficient was calculated for each variable group. Finally, only one variable with the highest correlation coefficient was selected from each variable group. Figure 11 shows the variables for Gangneung that were selected for each variable group, by applying the forward selection method and correlation analysis.

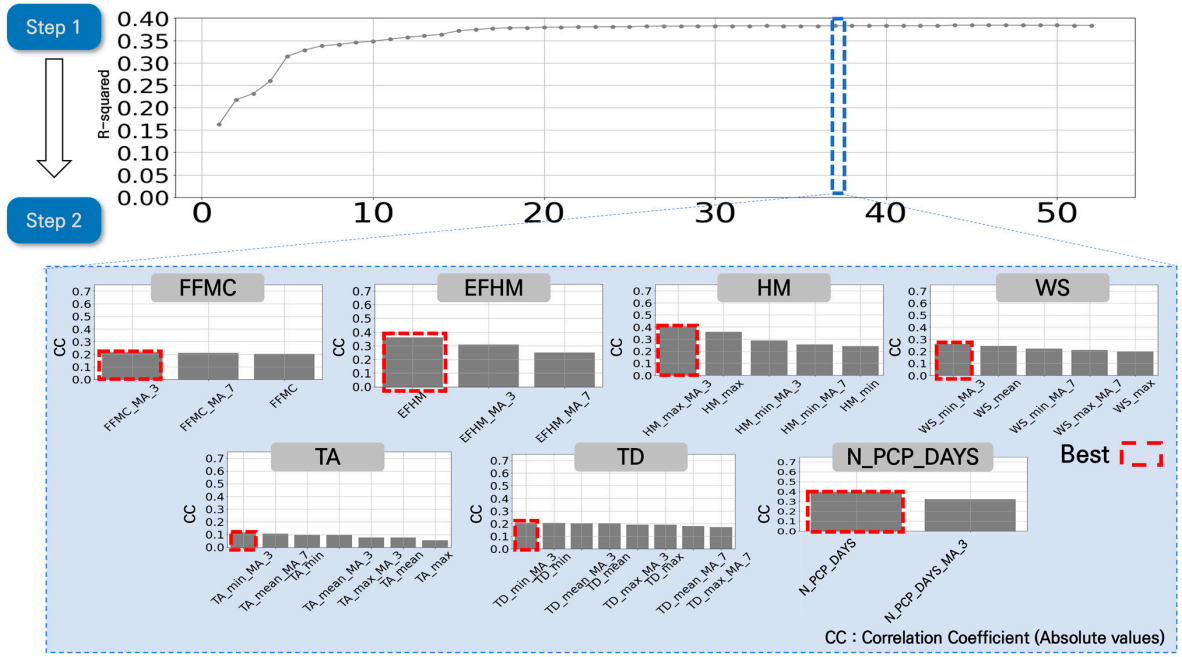


Figure 11. The process of selecting the variables using forward selection and correlation analysis (Gangneung). In Step 1, the variables are selected by applying the forward selection method. In Step 2, we created groups of variables with the same characteristics and analyzed the correlation between each group of variables and the dependent variable. After that, the most highly correlated variable was selected, one by one, from each group.

4.5. Multicollinearity Analysis

Multicollinearity is characterized by a strong correlation between the independent variables, which can affect a model's prediction accuracy [69]. Hence, it is necessary to identify whether multicollinearity exists between the independent variables. This study utilized the variance inflation factor (VIF) to prevent multicollinearity from undermining the model's predictive accuracy. Equation (15) represents the VIF equation, where R_i^2 is the regression coefficient obtained by excluding the i-th variable. Generally, the variable must be removed when the VIF exceeds 10 [24]. Therefore, this study uses VIF analysis to eliminate variables with VIFs > 10, in order to address the multicollinearity problem.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{15}$$

Figure 12 shows the VIF values of the selected variables by target area. In all the target areas, the VIFs of the dew point temperature and air temperature were greater than 10, with the dew point temperature having the highest VIF. Therefore, the dew point temperature was removed from all the target areas. The VIF analysis was then reapplied and all the variables had a VIF of 10 or less. Finally, the selected variables for each target area are shown in Table 3.

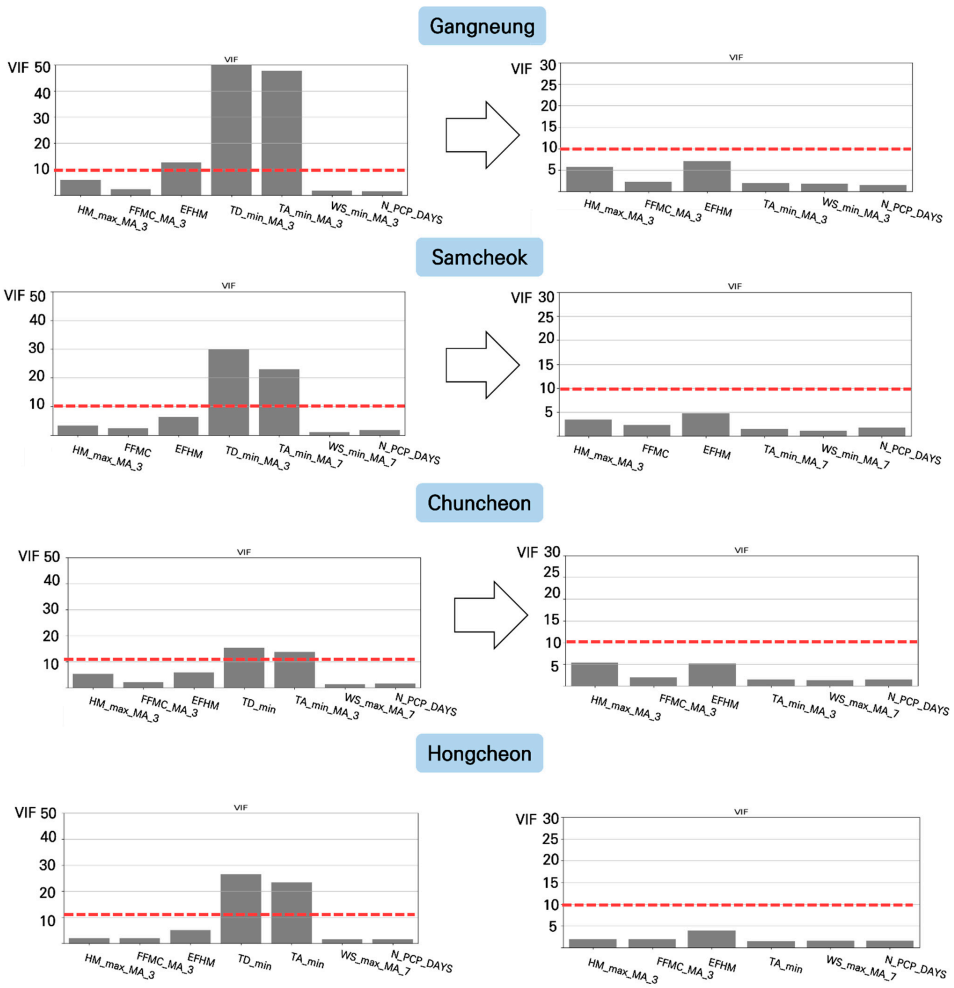


Figure 12. VIF analysis results on the selected variables by target area. Here, the red-dashed straight line represents VIF = 10. After removing the dew point temperature with the highest VIF and reanalyzing the data, we found that the VIF dropped below 10 in all areas.

Table 3. Final variables selected using the variable selection method and multicollinearity analysis for each target area. Where, 3 days means a 3-day moving average, and max and min mean maximum and minimum, respectively.

Region		Relative Humidity	FFMC	Effective Humidity	Temperature	Wind Speed	No Precipitation Days
Gangneung	MA	3 days	3 days	today	3 days	3 days	today
	Type	max	-	-	min	min	
Samcheok	MA	3 days	today	today	7 days	7 days	today
	Type	max	-	-	min	min	-
Chuncheon	MA	3 days	3 days	today	3 days	7 days	today
	Type	min	-	-	min	max	-
Hongcheon	MA	3 days	3 days	today	today	7 days	today
	Type	max	-	-	min	max	-

4.6. Synthetic Minority Oversampling Technique (SMOTE)

As shown in Table 2, all the target areas have class imbalance problems. Imbalanced data can adversely affect the prediction accuracy of machine learning models [59]. Imbalanced data can introduce bias into the model outcomes, due to the machine learning algorithm's predisposition to favor the majority class. This occurs because the algorithm typically learns to prioritize the most frequent class, neglecting the underrepresented class, which leads to poor predictive performance in terms of the less frequent categories. Additionally, imbalanced data can negatively impact the generalization ability of machine learning models. The limited representation of the minority class increases the risk of overfitting, as the model may memorize specific examples from the minority class rather than learning generalizable patterns, resulting in suboptimal predictions based on new data. One approach to address this issue is through sampling methods. There are two main sampling techniques, undersampling, where the data from the majority class are dropped to make the minority class more representative, and oversampling, where the data from the minority class are increased to make the majority class more representative. In the previous step, we removed the data from the non-forest fire days, while preserving the data on the forest fire days, as much as possible. This resulted in a clear difference in the distribution of the data according to the forest fire and non-forest fire days, but a class imbalance problem still existed. We needed a technique that could generate data that belonged to the distribution of forest fire data, rather than data that were randomly generated. Therefore, among the oversampling techniques that could satisfy the requirements of this study, we used the SMOTE, which has been used in several forest fire prediction studies [70–72]. SMOTE operates on the principle of the k-nearest neighbor (k-NN) algorithm. It identifies K-nearest neighbors based on data from the minority class and computes their differences on a straight line. These differences are then randomly multiplied by values between 0 and 1 to generate new data [73]. In this study, we generated the data so that the ratio of forest fire days and non-forest fire days was 1:2. Figure 13 shows the results of the SMOTE application, and a comparison of the results before and after the SMOTE application by target area is shown in Table 4.

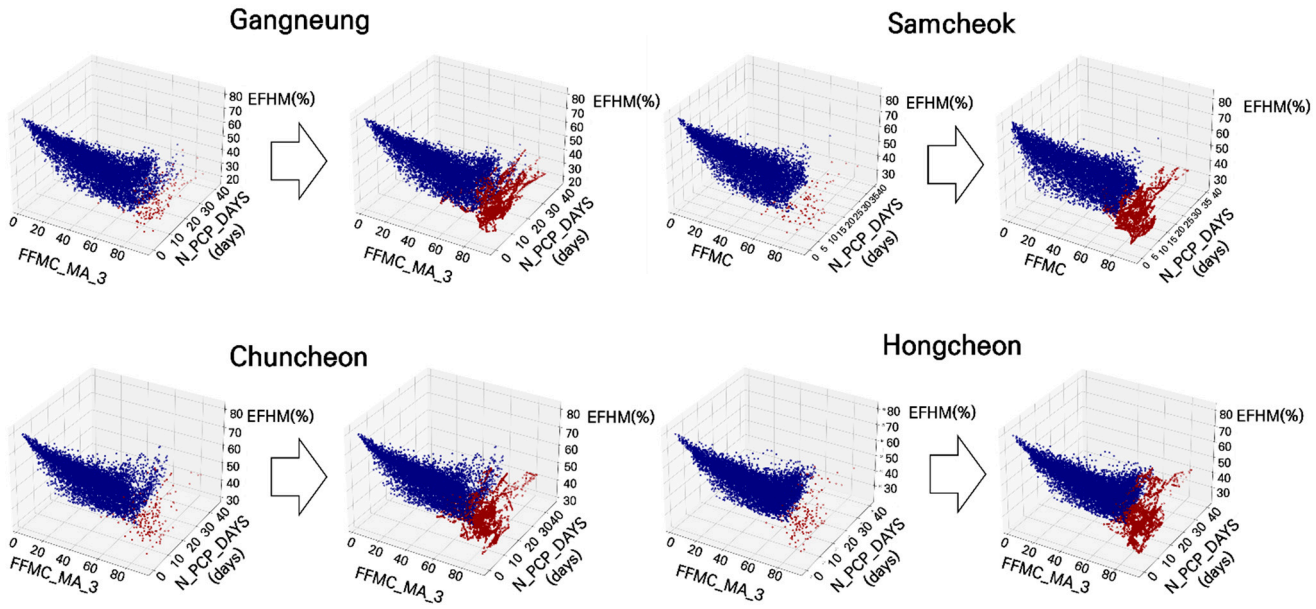


Figure 13. The results of applying the synthetic minority oversampling technique (SMOTE) are shown in these 3D scatter plots on the variables influencing forest fire occurrence, where the blue points represent non-forest fire conditions and the red points represent forest fire conditions. The SMOTE was applied to increase the number of red points (forest fire occurrences) and address the class imbalance.

Table 4. The results of oversampling using SMOTE by target area.

Region	Type	Forest Fire	Non-Forest Fire	Total
Gangneung	Before	163	6686	6849
	After	3343	6686	10,029
Samcheok	Before	99	6641	6740
	After	3320	6641	9961
Chuncheon	Before	186	5970	6156
	After	2985	5970	8955
Hongcheon	Before	171	6431	6602
	After	3215	6431	9646

5. Application and Results

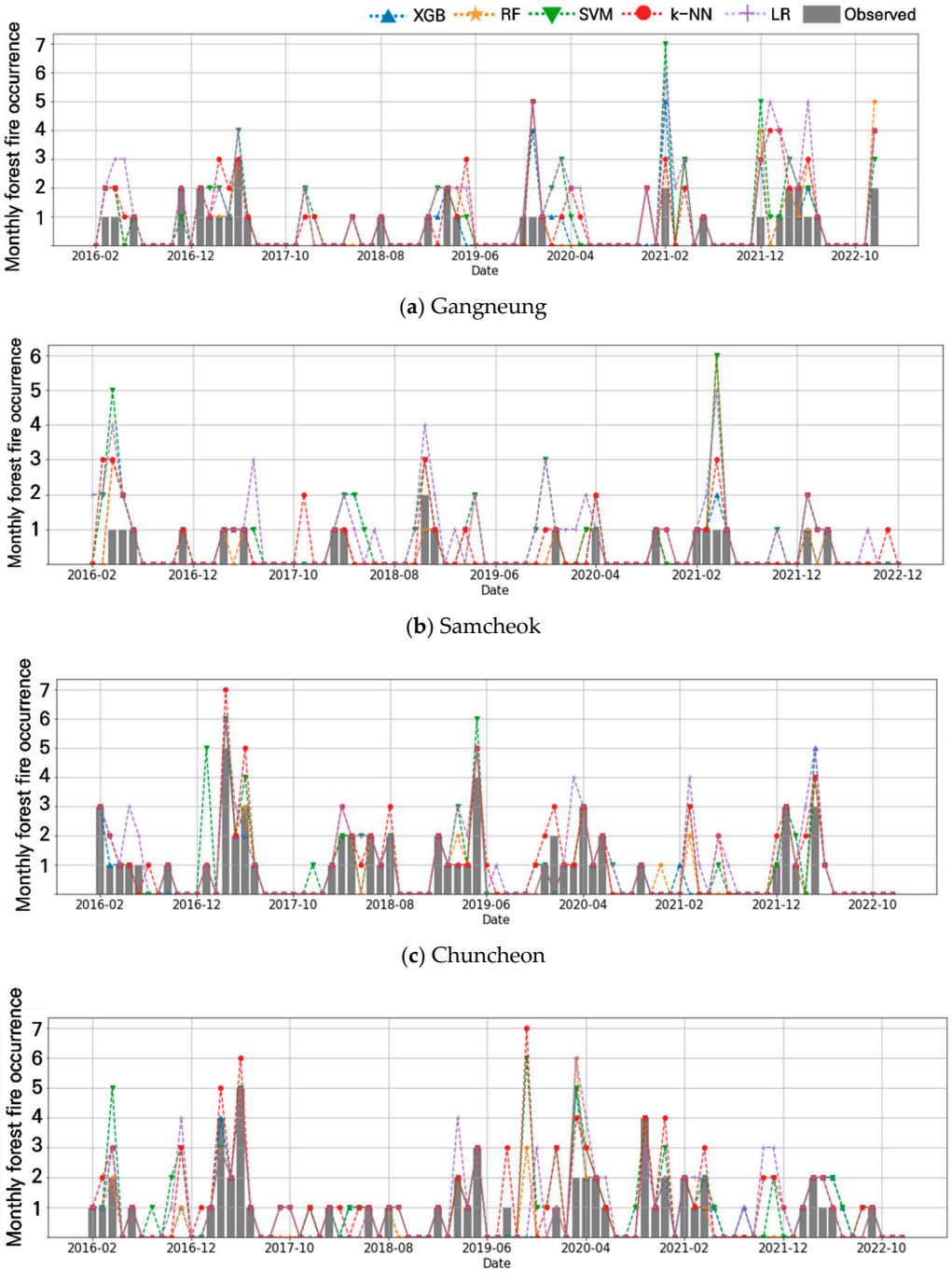
5.1. The Prediction Results from a Single Model

The multi-model ensemble-based forest fire prediction model proposed in this study comprised the five models mentioned above. Preceding the ensemble method's application, each model's prediction performance was assessed. All the models were trained using the data from 1991 to 2015 for training and the data from 2016 to 2022 for verification. Given the diverse units and ranges of the input data used for training, MinMax scaling was employed to standardize the data range. The formula for MinMax scaling is shown in Equation (16), where x is the input data, x_{min} is the minimum value, and x_{max} is the maximum value.

$$\frac{x - x_{min}}{x_{max} - x_{min}} \tag{16}$$

Figure 14 shows the forest fire prediction results from a single model and Table 5 shows the results of applying the evaluation metrics for each model. SVM and LR showed low prediction performance, because they predicted forest fire occurrence excessively compared to the other models in regard to all the target areas. The best performing models were XGB in regard to Yeongdong (Gangneung, Samcheok) and RF in regard to Yeongseo

(Chuncheon, Hongcheon), according to a comparison of the F1-score of the single models. A confusion matrix of the single models is summarized in Figure A1, in Appendix A. The FP was present in all the models. The FP of the best performing single model in regard to each target area was analyzed and the results are shown as a scatter plot in Figure 15. It can be seen that the FPs are located on the border of the TP and TN, and the data distribution of the FPs is closer to the data distribution of the TP than the TN. While FPs may indicate potential forest fire events if triggered by external factors (e.g., human activity), excessive FPs can diminish the model's reliability. Hence, there is a need to develop a model that accurately reflects actual forest fire occurrence data, while minimizing FPs to enhance the model's prediction performance and reliability.



(**a**) Gangneung

(**b**) Samcheok

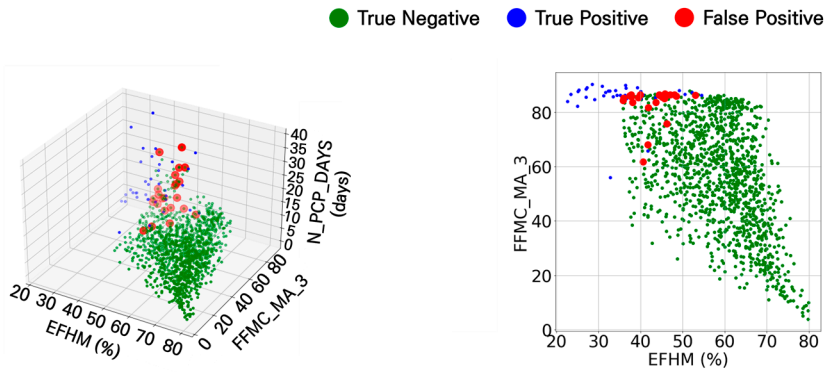
(**c**) Chuncheon

(**d**) Hongcheon

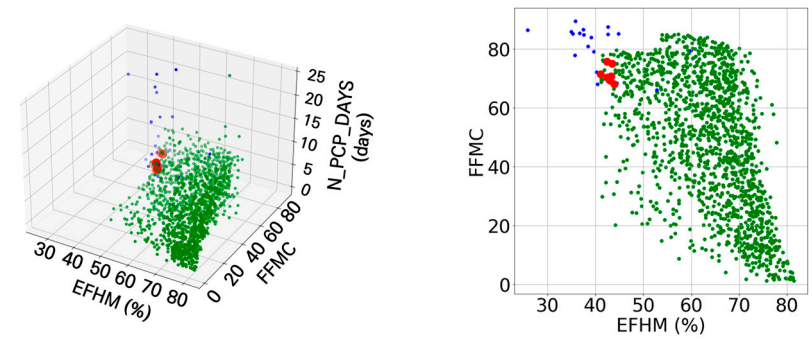
Figure 14. The prediction results for the validation period for each model.

Table 5. The evaluation results for each model.

Region	Type	Evaluation	XGB	SVM	RF	LR	k-NN	Ensemble
Gangneung	Training	F1-score	0.82	0.66	0.91	0.50	0.79	0.89
		Accuracy	0.99	0.97	0.99	0.95	0.98	0.99
	Validation	F1-score	0.73	0.61	0.72	0.54	0.64	0.78
		Accuracy	0.98	0.96	0.98	0.95	0.97	0.98
Samcheok	Training	F1-score	0.92	0.75	0.90	0.64	0.86	0.95
		Accuracy	0.99	0.99	0.99	0.98	0.99	0.99
	Validation	F1-score	0.72	0.56	0.69	0.45	0.67	0.78
		Accuracy	0.99	0.97	0.99	0.97	0.98	0.99
Chuncheon	Training	F1-score	0.88	0.71	0.93	0.55	0.81	0.93
		Accuracy	0.99	0.98	0.99	0.96	0.98	0.99
	Validation	F1-score	0.90	0.81	0.90	0.74	0.82	0.93
		Accuracy	0.99	0.97	0.98	0.96	0.97	0.99
Hongcheon	Training	F1-score	0.88	0.75	0.93	0.62	0.83	0.93
		Accuracy	0.99	0.98	0.99	0.97	0.99	0.99
	Validation	F1-score	0.86	0.71	0.89	0.72	0.74	0.93
		Accuracy	0.98	0.97	0.99	0.97	0.97	0.99

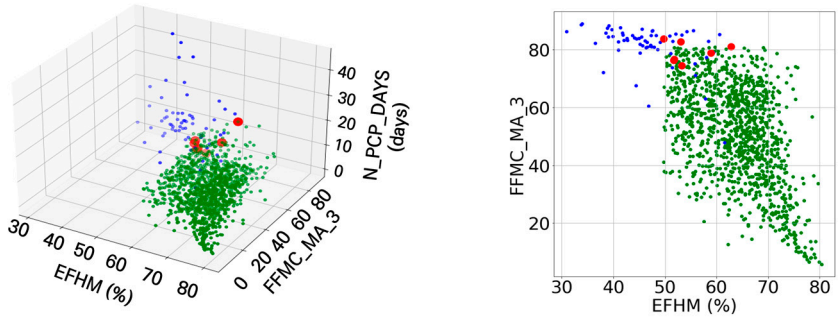


(**a**) Gangneung

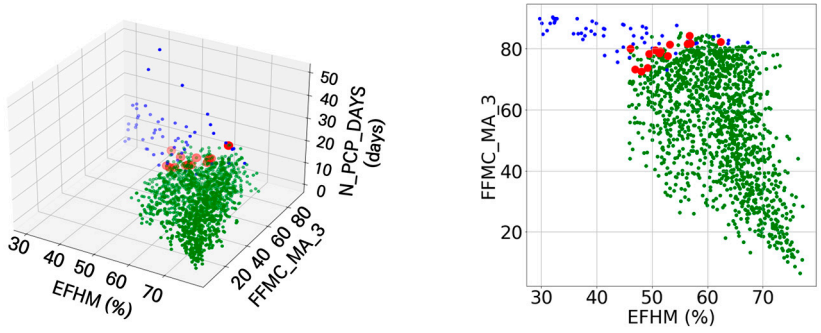


(**b**) Samcheok

Figure 15. *Cont.*



(**c**) Chuncheon

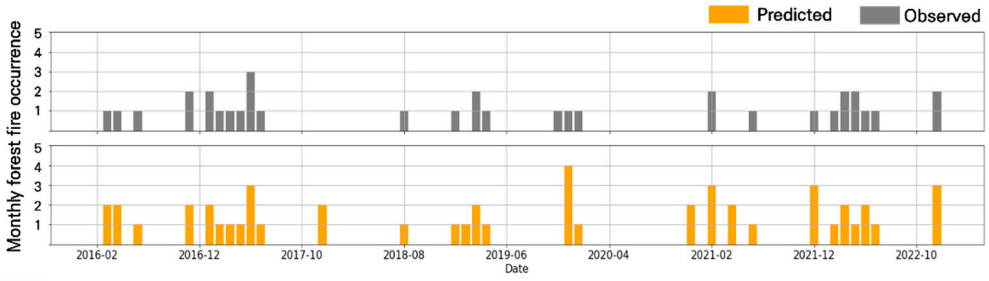


(**d**) Hongcheon

Figure 15. Three-dimensional and two-dimensional visualization of FP, TP, and TN, of the single best predictive model by the target area using FFMC, EFHM, and N_PCP_DAYS. Data points judged to be FPs were identified at the boundaries of the TP and the TN in all the target areas.

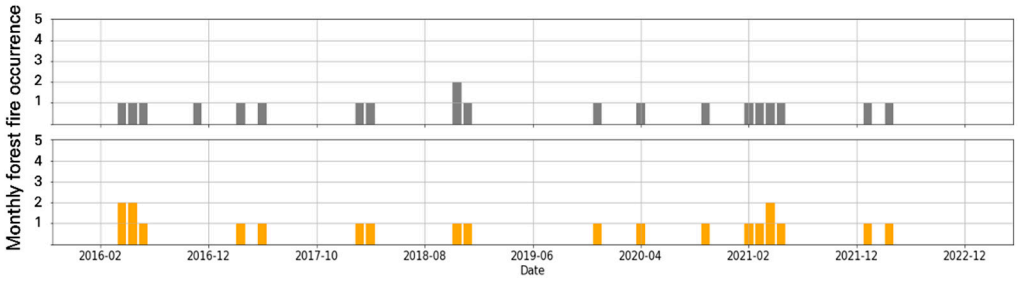
5.2. The Prediction Results from the Multi-Model Ensemble (MME) Method

The forest fire prediction model developed in this study employs the ensemble method to address the aforementioned limitations of single model predictions. If at least four of the five models predict a forest fire occurrence for given input data, the ensemble model predicts a forest fire event. The training period spanned 1991–2015, consistent with the single model training period, while the verification period extended from 2016 to 2022. F1-score and accuracy metrics were utilized to assess the ensemble model's prediction performance. Figure 16 shows the forecasting results from the MME model for the validation period, and the confusion matrix is shown in Figure A1, in Appendix A. We found that the MME model had better prediction results in regard to all the target areas. In Gangneung, the F1-score of the MME model was 0.78, which is about 6.8% higher than the best single model. The F1-scores for Samcheok, Chuncheon, and Hongcheon also outperformed the best single model by 8.3%, 3.3%, and 4.5%, respectively. Figure A1, in Appendix A, shows that the FPs also decreased in regard to each target area, decreasing by six, three, four, and five, respectively. The abovementioned results are summarized in Table 5.

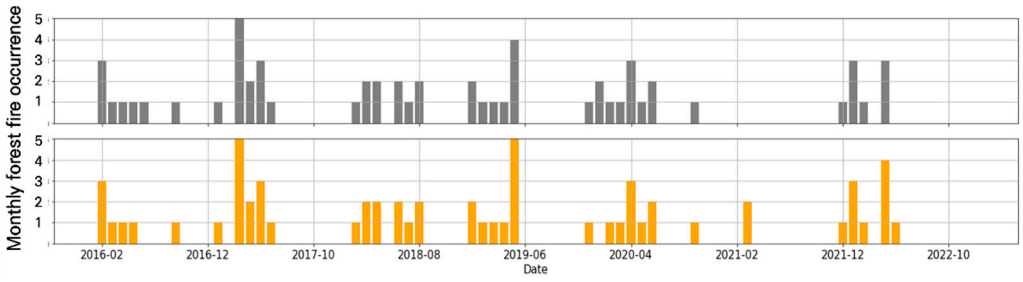


(**a**) Gangneung

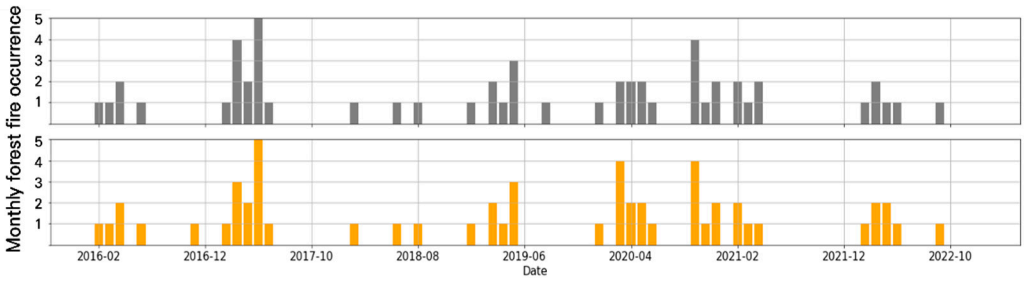
Figure 16. *Cont.*



(**b**) Samcheok



(**c**) Chuncheon



(**d**) Hongcheon

Figure 16. The prediction results for the validation period produced by the MME method.

6. Discussion

Forest fires in South Korea are predominantly caused by human activities, rather than natural factors like lightning. Over the past decade, 65% of forest fires occurred during the spring season. This suggests that forest fires are significantly influenced by meteorological conditions, in addition to anthropogenic factors. Therefore, developing a forest fire prediction model based on meteorological factors is crucial for effective prevention and early detection of forest fires. As some previous studies [20,23,32] have utilized the indices in the FWI System, this study utilized the FFMC, one of the indices in the FWI System, as an input variable for the machine learning model. It was determined that if the forest fire risk index was considered to be an important input variable in the forest fire prediction model, good prediction performance could be achieved. Most existing research on forest fire prediction models has focused on comparing the performance of various machine learning models to identify the best-performing model [21,22,24,74–76]. In contrast, this study applied ensemble techniques to combine the prediction results of multiple models, achieving superior prediction performance compared to using a single model.

Compared to similar studies in South Korea, ref. [25] conducted a deep learning-based forest fire prediction study for the spring season across the entire country, while this study advances the field by enabling more detailed predictions according to the season and geographical area. Similar to this study, ref. [74] applied machine learning, meteorological factors, and sampling techniques to predict forest fires; however, they divided Gangwon-do into nine zones instead of using municipal units, achieving an accuracy of 76.1%. In [76], machine learning, meteorological factors, and sampling techniques were also employed, as in this study, but forest fire prediction was conducted for the entire Gangwon-do province,

resulting in an accuracy of approximately 94%. The spatial scale of the target area in this study was set to the municipal unit, which likely contributed to the good level of prediction performance achieved.

In Table 5, the accuracy values are observed to be close to 1. This is because accuracy, as defined in Equation (9), is calculated by adding up the number of days with and without forest fire occurrences. To account for the class imbalance inherent in forest fire data, we also utilized the F1-score, which is a more appropriate metric for evaluating model performance in such unbalanced scenarios.

7. Conclusions

In this study, we developed a model to predict the number of forest fires by applying the MME technique to meteorological factors, which are the key variables affecting forest fires. To validate the model, we applied it to four regions in Gangwon-do, Korea, an area that has experienced large-scale forest fires. The main conclusions drawn from this study are as follows:

(1) When comparing the prediction results of a single model and the MME model using the F1-score, the MME model produced the best prediction results (Gangneung 6.8%, Samcheok 8.3%, Chuncheon 3.3%, and Hongcheon 4.5%). Additionally, the false positive (FP) rate decreased in all four target areas;

(2) Since the MME model developed in this study predicts the number of forest fires based on meteorological factors, combining it with meteorological forecast data could enable region-specific forest fire predictions, allowing for proactive measures to be implemented that would contribute to the preservation of forest resources and ecosystems;

(3) By providing predictions on the number of forest fires, intuitive information on how many fires are likely to occur can be delivered. This information can assist local forest fire managers during decision-making, when planning forest fire prevention strategies. Furthermore, if climate change scenario data are applied, it is possible to predict the number of future forest fires due to climate change and establish mid- to long-term forest fire prevention measures at the local level.

However, the forest fire prediction model developed in this study has limitations. Forest fires are also influenced by human and social factors, such as the tree species, tourism activity, and topographical features, but this study only accounted for meteorological factors. Future research should aim to improve the model by incorporating human and social factors. Additionally, the same meteorological variables were applied across all four seasons in this study. As the number of forest fires varies by season and the meteorological factors influencing forest fires may differ, future studies need to develop seasonal forest fire prediction models, by selecting the relevant meteorological factors for each season.

Author Contributions: Conceptualization, C.K. and B.K.; methodology, C.K.; validation, S.C.; formal analysis, S.C. and M.S.; investigation, S.C. and M.S.; writing—original draft preparation, S.C.; writing—review and editing, C.K. and B.K.; visualization, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant (2022-MOIS61-001) on the Development Risk Prediction Technology of Storm and Flood For Climate Change based on Artificial Intelligence funded by Ministry of Interior and Safety (MOIS, Republic of Korea).

Data Availability Statement: All data used during the study are included in this published article.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Appendix A

XGB (Train)		Predicted Values	
		N	Y
Observed Values	N	5354	50
	Y	4	123

SVM (Train)		Predicted Values	
		N	Y
Observed Values	N	5324	120
	Y	4	123

RF (Train)		Predicted Values	
		N	Y
Observed Values	N	5350	24
	Y	0	127

LR (Train)		Predicted Values	
		N	Y
Observed Values	N	5137	217
	Y	10	117

K-NN (Train)		Predicted Values	
		N	Y
Observed Values	N	5289	65
	Y	1	126

Ensemble (Train)		Predicted Values	
		N	Y
Observed Values	N	5333	21
	Y	6	121

XGB (Validation)		Predicted Values	
		N	Y
Observed Values	N	1309	23
	Y	2	34

SVM (Validation)		Predicted Values	
		N	Y
Observed Values	N	1292	40
	Y	2	34

RF (Validation)		Predicted Values	
		N	Y
Observed Values	N	1306	26
	Y	1	35

LR (Validation)		Predicted Values	
		N	Y
Observed Values	N	1275	57
	Y	1	35

K-NN (Validation)		Predicted Values	
		N	Y
Observed Values	N	1294	38
	Y	1	35

Ensemble (Validation)		Predicted Values	
		N	Y
Observed Values	N	1315	17
	Y	2	34

(**a**) Gangneung

XGB (Train)		Predicted Values	
		N	Y
Observed Values	N	5144	12
	Y	0	79

SVM (Train)		Predicted Values	
		N	Y
Observed Values	N	5104	52
	Y	0	79

RF (Train)		Predicted Values	
		N	Y
Observed Values	N	5140	16
	Y	0	79

LR (Train)		Predicted Values	
		N	Y
Observed Values	N	5140	16
	Y	0	79

K-NN (Train)		Predicted Values	
		N	Y
Observed Values	N	5132	24
	Y	1	78

Ensemble (Train)		Predicted Values	
		N	Y
Observed Values	N	5149	7
	Y	0	79

XGB (Validation)		Predicted Values	
		N	Y
Observed Values	N	1477	8
	Y	4	16

SVM (Validation)		Predicted Values	
		N	Y
Observed Values	N	1454	31
	Y	0	20

RF (Validation)		Predicted Values	
		N	Y
Observed Values	N	1473	12
	Y	3	17

LR (Validation)		Predicted Values	
		N	Y
Observed Values	N	1444	41
	Y	2	18

K-NN (Validation)		Predicted Values	
		N	Y
Observed Values	N	1466	19
	Y	0	20

Ensemble (Validation)		Predicted Values	
		N	Y
Observed Values	N	1480	5
	Y	4	16

(**b**) Samcheok

XGB (Train)		Predicted Values	
		N	Y
Observed Values	N	4770	24
	Y	6	118

SVM (Train)		Predicted Values	
		N	Y
Observed Values	N	4705	89
	Y	5	119

RF (Train)		Predicted Values	
		N	Y
Observed Values	N	4782	12
	Y	4	120

LR (Train)		Predicted Values	
		N	Y
Observed Values	N	4619	175
	Y	8	116

K-NN (Train)		Predicted Values	
		N	Y
Observed Values	N	4738	56
	Y	1	123

Ensemble (Train)		Predicted Values	
		N	Y
Observed Values	N	4784	10
	Y	7	117

XGB (Validation)		Predicted Values	
		N	Y
Observed Values	N	1169	7
	Y	5	57

SVM (Validation)		Predicted Values	
		N	Y
Observed Values	N	1153	23
	Y	3	59

RF (Validation)		Predicted Values	
		N	Y
Observed Values	N	1166	10
	Y	3	59

LR (Validation)		Predicted Values	
		N	Y
Observed Values	N	1142	34
	Y	5	57

K-NN (Validation)		Predicted Values	
		N	Y
Observed Values	N	1151	25
	Y	1	61

Ensemble (Validation)		Predicted Values	
		N	Y
Observed Values	N	1170	6
	Y	3	59

(**c**) Chuncheon

XGB (Train)		Predicted Values	
		N	Y
Observed Values	N	4987	26
	Y	3	112

SVM (Train)		Predicted Values	
		N	Y
Observed Values	N	4939	74
	Y	1	114

RF (Train)		Predicted Values	
		N	Y
Observed Values	N	4998	15
	Y	2	113

LR (Train)		Predicted Values	
		N	Y
Observed Values	N	4886	127
	Y	4	111

K-NN (Train)		Predicted Values	
		N	Y
Observed Values	N	4968	45
	Y	0	115

Ensemble (Train)		Predicted Values	
		N	Y
Observed Values	N	5000	13
	Y	3	112

XGB (Validation)		Predicted Values	
		N	Y
Observed Values	N	1404	14
	Y	3	53

SVM (Validation)		Predicted Values	
		N	Y
Observed Values	N	1379	39
	Y	3	53

RF (Validation)		Predicted Values	
		N	Y
Observed Values	N	1409	9
	Y	3	53

LR (Validation)		Predicted Values	
		N	Y
Observed Values	N	1385	33
	Y	0	56

K-NN (Validation)		Predicted Values	
		N	Y
Observed Values	N	1379	39
	Y	0	56

Ensemble (Validation)		Predicted Values	
		N	Y
Observed Values	N	1414	4
	Y	3	53

(**d**) Hongcheon

Figure A1. The confusion matrix.

References

1. 2020 Forest Basic Statistics. Available online: https://www.data.go.kr/data/15067764/fileData.do?recommendDataYn=Y (accessed on 12 October 2024).
2. Jang, E.; Kang, Y.; Im, J.; Lee, D.W.; Yoon, J.; Kim, S.K. Detection and Monitoring of Forest Fires Using Himawari-8 Geostationary Satellite Data in South Korea. *Remote Sens.* **2019**, *11*, 271. [CrossRef]
3. Lee, H.W.; Tak, S.H.; Lee, S.H. Numerical Experiment on the Variation of Atmospheric Circulation due to Wild Fire. *J. Environ. Sci. Int.* **2013**, *22*, 173–185. [CrossRef]
4. Park, H.S.; Lee, S.Y.; Chae, H.M.; Lee, W.K. A Study on the Development of Forest Fire Occurrence Probability Model using Canadian Forest Fire Weather Index -Occurrence of Forest Fire in Kangwon Province. *J. Korean Soc. Hazard Mitig.* **2009**, *9*, 50–100.
5. Ryu, S.R.; Choi, H.T.; Lim, J.H.; Lee, I.K.; Ahn, Y.S. Post-Fire Restoration Plan for Sustainable Forest Management in South Korea. *Forests* **2017**, *8*, 188. [CrossRef]
6. Jeong, K.O.; Kim, D.J. A Study on the Improvement of Safety Management by Analyzing the Current Status and Response System of Forest Fire Accidents. *J. Korean Soc. Disaster Inf.* **2022**, *18*, 457–469.
7. Bae, M.; Chae, H. Regional Characteristics of Forest Fire Occurrences in Korea from 1990 to 2018. *J. Korean Soc. Hazard Mitig.* **2019**, *19*, 305–313. [CrossRef]
8. Jeon, B.; Chae, H. A Study of Analysis on Relationship between Korea Forest Fire Occurrence and Weather Factor. *Korean Soc. Hazard Mitig.* **2017**, *17*, 197–206. [CrossRef]
9. Munger, T.T. Graphic Method of Representing and Comparing Drought Intensities.1. *Mon. Weather. Rev.* **1916**, *44*, 642–643. [CrossRef]
10. Westerling, A.L.; Hidalgo, H.G.; Cayan, D.R.; Swetnam, T.W. Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity. *Science* **2006**, *313*, 940–943. [CrossRef]
11. Flannigan, M.D.; Krawchuk, M.A.; de Groot, W.J.; Wotton, B.M.; Gowman, L.M. Implications of Changing Climate for Global Wildland Fire. *Int. J. Wildland Fire* **2009**, *18*, 483. [CrossRef]
12. Dowdy, A.J. Climatological Variability of Fire Weather in Australia. *J. Appl. Meteorol. Climatol.* **2018**, *57*, 221–234. [CrossRef]
13. Veraverbeke, S.; Rogers, B.M.; Goulden, M.L.; Jandt, R.R.; Miller, C.E.; Wiggins, E.B.; Randerson, J.T. Lightning as a Major Driver of Recent Large Fire Years in North American Boreal Forests. *Nat. Clim. Chang.* **2017**, *7*, 529–534. [CrossRef]
14. Dowdy, A.J.; Mills, G.A. Atmospheric and Fuel Moisture Characteristics Associated with Lightning-Attributed Fires. *J. Appl. Meteorol. Climatol.* **2012**, *51*, 2025–2037. [CrossRef]
15. WSL. Introduction to Fire. WikiFire. Available online: https://wikifire.wsl.ch/tiki-index515f.html?page=Introduction&structure=Fire (accessed on 14 October 2024).
16. Keetch, J.J.; Byram, G.M. *A Drought Index for Forest Fire Control*; Southeastern Forest Experiment Station: Asheville, NC, USA, 1968.
17. Burgan, R.E.; Cohen, J.D.; Deeming, J.E. *Manually Calculating Fire-Danger Rating—1978 National Fire-Danger Rating System*; USDA Forest Service: Washington, DC, USA, 1977.
18. Deeming, J.E.; Burgan, R.E.; Cohen, J.D. *The National Fire Danger Rating System—1978*; USDA Forest Service: Washington, DC, USA, 1977.
19. Nesterov, V.G. *Combustibility of the Forest and Methods for Its Determination*; USSR State Industry Press: Moscow, Russia, 1949.
20. Cortez, P.; Morais, A. *A Data Mining Approach to Predict Forest Fires Using Meteorological Data*; APPIA: Lisbon, Portugal, 2007.
21. Xie, Y.; Peng, M. Forest fire forecasting using ensemble learning approaches. *Neural Comput. Appl.* **2018**, *31*, 4541–4550. [CrossRef]
22. Lai, C.; Zeng, S.; Guo, W.; Liu, X.; Li, Y.; Liao, B. Forest Fire Prediction with Imbalanced Data Using a Deep Neural Network Method. *Forests* **2022**, *13*, 1129. [CrossRef]
23. Nebot, À.; Mugica, F. Forest Fire Forecasting Using Fuzzy Logic Models. *Forests* **2021**, *12*, 1005. [CrossRef]
24. Liang, H.; Zhang, M.; Wang, H. A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors. *IEEE Access* **2019**, *7*, 176746–176755. [CrossRef]
25. Sung, J.H.; Ryu, Y.; Seong, K.-W. Deep Learning-Based Prediction of Fire Occurrence with Hydroclimatic Condition and Drought Phase over South Korea. *KSCE J. Civ. Eng.* **2022**, *26*, 2002–2012. [CrossRef]
26. Carta, F.; Zidda, C.; Putzu, M.; Loru, D.; Anedda, M.; Giusto, D. Advancements in Forest Fire Prevention: A Comprehensive Survey. *Sensors* **2023**, *23*, 6635. [CrossRef]
27. Chen, B.; Bai, D.; Lin, H.; Jiao, W. FlameTransNet: Advancing Forest Flame Segmentation with Fusion and Augmentation Techniques. *Forests* **2023**, *14*, 1887. [CrossRef]
28. Peruzzi, G.; Pozzebon, A.; Van Der Meer, M. Fight Fire with Fire: Detecting Forest Fires with Embedded Machine Learning Models Dealing with Audio and Images on Low Power IoT Devices. *Sensors* **2023**, *23*, 783. [CrossRef] [PubMed]
29. Talaat, F.; ZainEldin, H. An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Comput. Appl.* **2023**, *35*, 20939–20954. [CrossRef]
30. Duangsuwan, S.; Klubsuwan, K. Accuracy Assessment of Drone Real-Time Open Burning Imagery Detection for Early Wildfire Surveillance. *Forests* **2023**, *14*, 1852. [CrossRef]

31. Zheng, S.; Zou, X.; Gao, P.; Zhang, Q.; Hu, F.; Zhou, Y.; Wu, Z.; Wang, W.; Chen, S. A Forest Fire Recognition Method Based on Modified Deep CNN Model. *Forests* **2024**, *15*, 111. [CrossRef]
32. Jesús, S.; Schulte, E.; Schmuck, G.; Camia, A.; Strobl, P.; Liberta, G.; Giovando, C.; Boca, R.; Sedano, F.; Kempeneers, P.; et al. Comprehensive Monitoring of Wildfires in Europe: The European Forest Fire Information System (EFFIS). In *Approaches to Managing Disaster-Assessing Hazards, Emergencies and Disaster Impacts*; IntechOpen: London, UK, 2012. [CrossRef]
33. Wahyono; Harjoko, A.; Dharmawan, A.; Adhinata, F.D.; Kosala, G.; Jo, K.-H. Real-Time Forest Fire Detection Framework Based on Artificial Intelligence Using Color Probability Model and Motion Feature Analysis. *Fire* **2022**, *5*, 23. [CrossRef]
34. Foggia, P.; Saggese, A.; Vento, M. Real-Time Fire Detection for Video-Surveillance Applications Using a Combination of Experts Based on Color, Shape, and Motion. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 1545–1556. [CrossRef]
35. Sharma, S.; Khanal, P. Forest Fire Prediction: A Spatial Machine Learning and Neural Network Approach. *Fire* **2024**, *7*, 205. [CrossRef]
36. Ahmad, F.; Waseem, Z.; Ahmad, M.; Ansari, M.Z. Forest Fire Prediction Using Machine Learning Techniques. In Proceedings of the 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), New Delhi, India, 1–3 May 2023; Volume 63, pp. 705–708. [CrossRef]
37. Cardil, A.; Monedero, S.; Ramírez, J.; Silva, C.A. Assessing and reinitializing wildland fire simulations through satellite active fire data. *J. Environ. Manag.* **2019**, *231*, 996–1003. [CrossRef]
38. Son, M.-W.; Kim, C.-G.; Kim, B.-S. Development of an Algorithm for Assessing the Scope of Large Forest Fire Using VIIRS-Based Data and Machine Learning. *Remote Sens.* **2024**, *16*, 2667. [CrossRef]
39. Shadrin, D.; Illarionova, S.; Gubanov, F.; Evteeva, K.; Mironenko, M.; Levchunets, I.; Belousov, R.; Burnaev, E. Wildfire Spreading Prediction Using Multimodal Data and Deep Neural Network Approach. *Sci. Rep.* **2024**, *14*, 2606. [CrossRef]
40. Marjani, M.; Mahdianpari, M.; Mohammadimanesh, F. CNN-BiLSTM: A Novel Deep Learning Model for Near-Real-Time Daily Wildfire Spread Prediction. *Remote Sens.* **2024**, *16*, 1467. [CrossRef]
41. Chen, T.; Guestrin, C. XGBoost A Scalable Tree Boosting System. In Proceedings of the KDD'16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
42. Narvaez, J.P.; Guillen, M.; Alcañiz, M. Prediction Motor Insurance Claims Using Telematics Data-XGBoost versus Logistic Regression. *Risks* **2019**, *7*, 70. [CrossRef]
43. Ma, M.; Zhao, G.; He, B.; Li, Q.; Dong, H.; Wang, S.; Wang, Z. XGBoost-based method for flash flood risk assessment. *J. Hydrol.* **2021**, *598*, 126382. [CrossRef]
44. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
45. Taylor, S.W.; Alexander, M.E. Science, technology, and human factors in fire danger rating: The Canadian experience. *Int. J. Wildland Fire* **2006**, *15*, 121–135. [CrossRef]
46. Ma, W.; Feng, Z.; Cheng, Z.; Chen, S.; Wang, F. Identifying Forest Fire Driving Factors and Related Impacts in China Using Random Forest Algorithm. *Forests* **2020**, *11*, 507. [CrossRef]
47. Siroky, D.S. Navigating Random Forests and related advances in algorithmic modeling. *Stat. Surv.* **2009**, *3*, 147–163. [CrossRef]
48. Park, S.W.; Kim, C.G.; Youm, S.K. Establishment of an IoT-based smart factory and data analysis model for the quality management of SMEs die-casting companies in Korea. *Int. J. Distrib. Sens. Netw.* **2019**, *15*, 10. [CrossRef]
49. Rosadi, D.; Andriyani, W.; Arisanty, D.; Agustina, D. Prediction of forest fire occurrence in peatlands using machine learning approaches. In Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, Yogyakarta, Indonesia, 10–11 December 2020; pp. 48–51. [CrossRef]
50. Yue, W.; Ren, C.; Liang, Y.; Liang, J.; Lin, X.; Yin, A.; Wei, Z. Assessment of wildfire susceptibility and wildfire threats to ecological environment and urban development based on GIS and multi-source data: A case study of Guilin, China. *Remote Sens.* **2023**, *15*, 2659. [CrossRef]
51. 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]
52. Rezaei Barzani, A.; Pahlavani, P.; Ghorbanzadeh, O. Ensembling of decision trees, KNN, and logistic regression with soft-voting method for wildfire susceptibility mapping. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *10*, 647–652. [CrossRef]
53. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. KNN model-based approach in classification. In *On the Move to Meaningful Internet Systems 2003*; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2003, pp. 986–996. [CrossRef]
54. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
55. Park, J.J.; Kim, K.M. Recognition of Handwritten Numerals using SVM Classifiers. *J. Korea Inst. Converg. Signal Process.* **2007**, *8*, 136–142.
56. Gigović, L.; Pourghasemi, H.R.; Drobnjak, S.; Bai, S. Testing a New Ensemble Model Based on SVM and Random Forest in Forest Fire Susceptibility Assessment and Its Mapping in Serbia's Tara National Park. *Forests* **2019**, *10*, 408. [CrossRef]
57. 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]
58. Han, S.; Qubo, C.; Meng, H. Parameter selection in SVM with RBF kernel function. In Proceedings of the World Automation Congress 2012, Puerto Vallarta, Mexico, 24–28 June 2012; pp. 1–4.

59. Elrahman, S.M.; Abraham, A. A Review of Class Imbalance Problem. *J. Netw. Innov. Comput.* **2013**, *1*, 9.
60. Korea Forest Service. 2023 Forest Cover Map (1:5000). Available online: https://map.forest.go.kr/forest/ (accessed on 13 October 2024).
61. Kim, N.; Kwak, J.; Kim, M.I. Numerical Simulations of Large Forest Fires in the East Coastal Region of Korea Considering the Yangganjipung Local Wind. *J. Korean Soc. Hazard Mitig.* **2021**, *21*, 39–48. [CrossRef]
62. GloH2O. Köppen-Geiger Climate Classification. Available online: https://www.gloh2o.org/koppen/ (accessed on 13 October 2024).
63. Won, M.S.; Lee, M.B.; Lee, W.K.; Yoon, S.H. Prediction of Forest Fire Danger Rating over the Korean Peninsula with the Digital Forecast Data and Daily Weather Index (DWI) Model. *Korean J. Agric. For. Meteorol.* **2021**, *14*, 1–10. [CrossRef]
64. Choi, J.; Park, S. Impacts of Seasonal Precipitation and Dryness on Regional Occurrences of Wildfires in South Korea. *J. Korean Assoc. Reg. Geogr.* **2020**, *26*, 307–319. [CrossRef]
65. Jeon, H.-E.; Ha, K.-J.; Kim, H.-R. A Study on Characteristics of Climate Variability and Changes in Weather Indexes in Busan Since 1904. *Atmosphere* **2023**, *33*, 1–20. [CrossRef]
66. Van Wagner, C.E. Development and structure of the Canadian forest fire weather index system. In *Forestry Technical Report 2014 Canadian Forestry Service*; CABI: Wallingford, UK, 1987; p. 35.
67. Schwertman, N.C.; Owens, M.A.; Adnan, R. A simple more general boxplot method for identifying outliers. *Comput. Stat. Data Anal.* **2004**, *47*, 165–174. [CrossRef]
68. Reif, M.; Shafait, F. Efficient feature size reduction via predictive forward selection. *Pattern Recognit.* **2014**, *47*, 1664–1673. [CrossRef]
69. Pourghasemi, H.R.; Gayen, A.; Lasaponara, R.; Tiefenbacher, J.P. Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling. *Environ. Res.* **2020**, *184*, 109321. [CrossRef] [PubMed]
70. Pérez-Porras, F.-J.; Triviño-Tarradas, P.; Cima-Rodríguez, C.; Meroño-de-Larriva, J.-E.; García-Ferrer, A.; Mesas-Carrascosa, F.-J. Machine Learning Methods and Synthetic Data Generation to Predict Large Wildfires. *Sensors* **2021**, *21*, 3694. [CrossRef] [PubMed]
71. Al-Bashiti, M.K.; Naser, M.Z. Machine learning for wildfire classification: Exploring blackbox, eXplainable, symbolic, and SMOTE methods. *Nat. Hazards Res.* **2022**, *2*, 154–165. [CrossRef]
72. Xu, Y.; Zhou, K.; Zhang, F. Modeling Wildfire Initial Attack Success Rate Based on Machine Learning in Liangshan, China. *Forests* **2023**, *14*, 740. [CrossRef]
73. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
74. Chae, K.; Lee, Y.; Cho, Y.; Park, J. Development of a Gangwon Province Forest Fire Prediction Model using Machine Learning and Sampling. *Korea J. BigData* **2018**, *3*, 71–78. [CrossRef]
75. Stojanova, D.; Kobler, A.; Ogrinc, P.; Ženko, B.; Džeoski, S. Estimating the risk of fire outbreaks in the natural environment. *Data Min. Knowl. Discov.* **2012**, *24*, 411–442. [CrossRef]
76. Lee, C.; Lim, M.; Lee, Y.M. Machine Learning for Big Data Analytics in Development of Wildfire Prediction Models. *J. Korean Soc. Hazard Mitig.* **2023**, *23*, 29–39. [CrossRef]