Hindawi
Advances in Meteorology
Volume 2022, Article ID 1127628, 12 pages
https://doi.org/10.1155/2022/1127628

Research Article

Frost Forecasting considering Geographical Characteristics

Hyojeoung Kim,[1] **Jong-Min Kim**,[2] **and Sahm Kim**[1]

[1]*Department of Applied Statistics, Chung-ang University, Seoul, Republic of Korea*
[2]*Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, USA*

Correspondence should be addressed to Sahm Kim; sahm@cau.ac.kr

Received 22 June 2022; Accepted 12 September 2022; Published 25 September 2022

Academic Editor: Yaolin Lin

Regional accuracy was examined using extreme gradient boosting (XGBoost) to improve frost prediction accuracy, and accuracy differences by region were found. When the points were divided into two groups with weather variables, Group 1 had a coastal climate with a high minimum temperature, humidity, and wind speed and Group 2 exhibited relatively inland climate characteristics. We calculated the accuracy in the two groups and found that the precision and recall scores in coastal areas (Group 1) were significantly lower than those in the inland areas (Group 2). Geographic elements (distance from the nearest coast and height) were added as variables to improve accuracy. In addition, considering the continuity of frost occurrence, the method of reflecting the frost occurrence of the previous day as a variable and the synthetic minority oversampling technique (SMOTE) pretreatment were used to increase the learning ability.

1. Introduction

While the recent rise in global average temperatures has accelerated the flowering of crops, sudden cold waves have occurred due to increased temperature volatility, causing crop damage. When a crop experiences a low temperature, the tissue freezes, causing the cell or chloroplast membrane to harden and be destroyed or the cells to dry and die. Frost, which causes direct damage to crops, refers to small ice crystals frozen on the ground or objects. When moist air contacts a cold surface at a temperature below the dew point of water vapor in the air, condensation occurs, and frost begins to form if the surface temperature is below the freezing temperature of water. Damage caused by frost has recently occurred worldwide, and interest in frost prediction has been increasing [1].

In July 2021, about 30% of local coffee trees were damaged by sudden subzero weather and frost in the Minas Gerais state of Brazil, the world's largest coffee producer (CONAB, 2021). Coffee prices surged nearly 13% in response to the frosts to a 6–1/2-year high. In addition, severely damaged farms take three years to recover their crops, which is expected to cause substantial long-term damage [2]. France suffered from spring frost for the second consecutive year, recording the coldest April ever in France in 22 years, following April 2021. The total damage was $2 billion due to massive frost. About 80% of vineyards were damaged, and wine production decreased by about 27% year-over-year in 2021 [3]. In 2022, temperatures in northern France fell to −9°C, reproducing the fierce cold of the midwinter, which is expected to significantly influence grapes and fruit trees, such as peach and apricot. As frost may have serious consequences on crop production, so actions must be taken to minimize damaging effects, and studies on frost occurrence and prediction have also been published steadily.

Using Stevenson screen temperature thresholds of 2°C or below as an indicator of frost at the ground level, Crimp et al. [4] demonstrated that, across southern Australia, despite a warming trend of 0.17°C per decade since 1960, the 'frost season' length has increased, on average, by 26 days across the southern portion of Australia compared with the long-term mean from 1960 to 1990. Unlike the recent growth of plants, which has accelerated due to unseasonably warm weather due to global warming, large temperature fluctuations, such as sudden cold waves, are causing hundreds of thousands of hectares of damage every year in conjunction

with the vulnerability to cold waves of plants that have grown rapidly due to warming. The damage caused by temperature volatility is increasing, and frost prediction is also becoming more difficult. Accordingly, attempts to accurately predict frost occurrence are ongoing.

Ghielmi and Eccel [5] demonstrated that the multilayer perceptron (MLP) model captures the specific features of the climatic conditions of any site to predict frost, even if the neural network does not represent functional relationships with nocturnal cooling in explicit form. In addition, Sallis et al. [6] studied the dependencies (correlation) between temperature, humidity, wind speed, precipitation, and barometric pressure variables using SOM. Studies have found that the wind direction during a frost is primarily influenced by the mountains (the Los Andes Mountains) in the valleys of the O'Higgins region of Chile; therefore, an analysis of the geospatial factors is necessary.

Lee et al. [7] selected the decision tree as a frost prediction model because it has a higher probability of detection than the logic regression analysis with data from six observatories on the Korean Peninsula. In addition, using autoregressive models with external input and MLP models, Castaneda-Miranda and Castano [8] predicted the temperature inside a greenhouse using the external air temperature, ambient air relative humidity, wind speed, global solar radiation flux, and internal air relative humidity variables. The R^2 of MLP exhibited higher performance in summer and winter at 0.9549 and 0.9590, respectively. In addition, Diniz et al. [9] generated possible local-scale predictors of frost occurrence, which included longitude, latitude, elevation, relative altitude, relief orientation, and Euclidean distance from hydrography. Three machine learning classifiers (random forest (RF), support vector machine (SVM), and MLP) were compared in order to determine which would most accurately predict frost occurrence, and RF has been found to be the most proficient algorithm.

Zendehboudi and Li [10] predicted the frost thickness and density on vertical and horizontal cryogenic surfaces. The hybrid adaptive neuro-fuzzy inference system, least-square support vector machine (SVM) with the genetic algorithm, radial basis function neural network with the genetic algorithm, and MLP models were compared. In all four cases, R^2 in the MLP model was about 0.9994, 0.9997, 0.9953, and 0.9965 for the frost thickness and density on horizontal and parallel surfaces, respectively, exhibiting the best performance. Additionally, Diedrichs et al. [11] demonstrated improved model performance by increasing the reproducibility in the RF and logistic regression models when the synthetic minority oversampling technique (SMOTE) was applied.

In addition, Rostamian and Halabian [12] investigated the probability and frequency of frost days using the Markov chain model. Two-day continuities in all stations revealed the minimum return period. All analyzed stations in the studied area, except for Nehbandan, which generally does not experience frost days, were characterized using the first-order Markov chain, indicating that frost days depend on past weather conditions.

Ding et al. [13] predicted the possibility of future frost with an SVM model using the historical values of temperature, humidity, and radiation. Temperature is a key factor in frost prediction models, with humidity helping generate an early warning for a relatively long period, such as within 2 or 3 h. Further, radiation has demonstrated improved sensitivity by reflecting changes in some areas in a short period. Finally, when predicting the next 1, 2, and 3 h of frost with the SVM model, the reproduction rates were 100%, 99.3%, and 99.8%, respectively.

Tamura et al. [14] compared the SVM results using the simple moving average and exponential moving average as the past values of the temperature and vapor pressure variables. Models using exponential moving averages perform a few percentage points better than models using simple moving averages in terms of the F1-score (the harmonic mean of the precision or recall) measurements.

Rozante et al. [15] corrected the frost index (IG) by correcting the weight of the variable numerically calculated by the local weather forecast model. The weight was adjusted, so that the temperature had the greatest contribution, followed by pressure and wind, and the other variables were determined with the constraint that the weight sum was 1.

Wassan et al. [16] predicted frost with a convolutional neural network model. For the one-dimensional data analysis, 1D convolution was used, and the accuracy was 97.6% for 30,000 repetitions and 98.6% for 50,000 repetitions.

In previous work [17], we compared the results of the SVM, RF, and MLP models, which were frequently mentioned in other papers, and the extreme gradient boosting (XGBoost) models, which have recently been frequently used in various fields. Daily statistics (total, average, maximum, minimum, etc.) were calculated with weather factors, such as wind speed, temperature, humidity, precipitation, and clouds, using the frost history and ground observation data for 20 years at 53 domestic points in Korea. Using XGBoost, SVM, RF, and MLP models, various hyperparameters were applied as training data to select the best model for each model, and the final model performance was evaluated from the testing data using various model evaluation criteria, such as the accuracy, F1-score, and critical success index (CSI). Compared to other models, XGBoost performed best with 90.4% accuracy and 64.4% on the CSI, followed by SVM with 89.7% accuracy and 61.2% on the CSI. The RF and MLP models performed similarly with about 89% accuracy and 60% on the CSI. The model was compared only as a weather variable, confirming that XGBoost had the best performance. However, the performance varies greatly from branch to branch. In this study, k -means clusters were used to increase the accuracy of the frost prediction model for each observation point. The observation points were clustered by k -means clustering, and we were able to find the differences and characteristics of the clustered groups through this. k-means clustering was conducted with weather variables, and the characteristics were examined according to the group, confirming that most islands were distributed into one group. Accordingly, frost was predicted by adding the shoreline distance and altitude, the geographical characteristics reflecting the terrain features. The

accuracy performance was improved from 90.4% to 90.8%, and CSI improved from 64.5% to 65.8%.

Comparing the model only as a weather variable confirmed that XGBoost performed the best. In this study, to improve the accuracy of the frost prediction model, a multipronged approach was attempted to determine other important variables for frost prediction.

In addition, the difference in accuracy was different depending on the difference in the frequency of frost occurrence by cluster. Accordingly, SMOTE was conducted to increase the learning rate, but there was no significant effect on accuracy improvement. However, considering that frost is continuous depending on the station, the accuracy increased by adding geographic characteristics and a categorical variable indicating whether frost occurred the previous day. Compared to the previous model (with added terrain features), accuracy improved from 90.8% to 92.6%, and CSI improved from 65.8% to 71.4%.

This paper is organized as follows. Section 2 introduces the various methods used in the study. Section 3 explains the analysis data and process of adding variables by clustering, and Section 4 concludes with a summary of the research results.

2. Methods

This section introduces the various methods used in the study. The XGBoost model, which performed best in previous studies, was used to predict frost occurrence. In addition, 53 stations were divided into Groups 1 and 2 using the k-means clustering method to compare accuracy.

2.1. Extreme Gradient Boosting. The XGBoost model is a decision tree-based algorithm that improves gradient boosting and is used in various studies [18]. Gradient boosting is a machine learning technique that increases predictive power by sequentially generating a model by supplementing the predictive error of the previous tree with the slope-lowering method using gradient descent. Through the repetition process of creating a new prediction model by focusing on poorly predicted individuals, a strong model is generated through a combination of several weak models. The XGBoost model consists of M decision trees, as in the following expression, where f denotes one decision tree, and F denotes a function of all decision trees:

$$Y_i = \sum_{m=1}^{M} f_m(x_i), \quad f_m \in F. \tag{1}$$

In the regression process, the model is expressed by the following equation:

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, Y_i) + \sum_{m=1}^{M} \Omega(f_m), \quad \theta = (f_1, f_2, \ldots, f_m)e, \tag{2}$$

where l represents the loss function, and Ω indicates the regulation to prevent overfitting. The regulation equation is as follows:

$$\Omega(f) = \frac{\gamma T + 1}{2\lambda \sum_{j=1}^{T} \omega_j^2}, \tag{3}$$

where T denotes the number of nodes in the decision tree, Ω represents a weight vector, and γ and λ are penalty elements.

2.2. K-Means Clustering. The k-means clustering algorithm is a method of dividing given data into several groups. In this case, the groups are divided by minimizing the cost functions, such as the distance-based intergroup dissimilarity, where the similarity between data objects in the same group increases, and the similarity with data objects in different groups decreases.

Choosing the best value of k in the various k-means algorithms can be difficult [19]. In this study, a silhouette score was used to determine the k value of the k-means algorithm. The silhouette coefficient is calculated by considering the mean intracluster distance a and the mean nearest-cluster distance b for each data point [20]. The value of the silhouette coefficient $s(i)$ for the i^{th} $x(i)$ is defined by the following equation:

$$s(i) = \frac{b^{(i)} - a^{(i)}}{\max\left(a^{(i)}, b^{(i)}\right)}, \tag{4}$$

where $a(i)$ represents data cohesion in a cluster and is the average distance from the rest of the data in the same cluster as $x(i)$. A smaller distance indicates higher cohesion. In addition, $b(i)$ represents intercluster separation and is the average distance from $x(i)$ and all data in the closest cluster. To optimize the number of clusters, $b(i)$ is large, $a(i)$ is small, and $s(i)$ is close to 1.

When the k value is determined, the k-means algorithm randomly specifies k centroids in the dataset, and each data point is allocated as a group of nearest centroids. In the assigned group, the process is repeated until the centroids converge by reassigning them. The group is classified as the side closest to the convergent final centroids.

3. Data Processing and Results

3.1. Data Preprocessing Process. Frost occurs when the dew point temperature of the air near the surface is below the freezing point; thus, the occurrence of frost was predicted for 24 h (from noon of the base date to noon of the next day) based on the continuity of the temperature. Data were extracted from October to April between 2000 and 2021 when frost occurs due to the influence of the changing of seasons. In addition, data from the point when no frost occurred from October of the base year to April of the following year were removed. During the removal process, frost data were extracted from 53 observation points in Korea, and weather data (temperature, wind speed, humidity, cloud, precipitation, and solar radiation) were collected for each observation, as listed in Table 1. The statistics (average, total sum, maximum, minimum, and standard variance) were calculated from noon of the reference date to noon of the next day, as listed in Table 2, to improve the accuracy of the analysis and obtain meaningful information

Table 1: Description of weather variables.

Weather variable	Description	Observation period and unit
Temperature	Temperature is measured from 1.2 to 1.5 meters above the ground. Automatically observed by a platinum resistance thermometer.	Observation period: 1 hour, unit: ˚C
Wind speed	The wind speed depends on the height from the ground, so it means the average wind speed at a height of 10 m above the ground. Automatically observed using wind speed sensors such as optical chopper or ultrasonic.	Observation period: 1 hour, unit: m/s
Humidity	Humidity is expressed as relative humidity. (The ratio of the amount of water vapor actually contained in the atmosphere to the maximum amount of water vapor that can be contained by the temperature at that time). Automatically observed by capacitive hygrometer.	Observation period: 1 hour, unit: %
Cloudiness	The total cloud is the 10th fraction of the sky that all clouds cover. Visually observed manually.	Observation period: 1 hour, unit: 1/10
Precipitation	Precipitation is measured liquid precipitation such as rain and dew rain or or by melting solid precipitation, such as snow or hail. Mainly automatically observed with conductive precipitators or weight-type precipitators.	Observation period: 1 hour, unit: mm
Solar radiation	Short wave radiation or insolation of solar radiation with a wavelength of 0.29 to 3.0 μm. Automatically observed by the Solar System	Observation period: 1 hour, unit: MJ/m^2
Frost	Frost is a phenomenon in which ice crystals are attached to the ground or an object on the ground by sublimation. Visually observed manually.	Observation period: 1 hour, unit: 0 or 1

Table 2: Derived variables.

Period	Weather factor	Descriptive statistics
For 1 day (from noon the day before to noon the next day)	Temperature Humidity Wind speed Cloudiness Precipitation	Min, max, difference (max-min), mean, sum, standard deviation
For 3 days (from 3 days prior to time of occurrence)	Precipitation	Mean, sum
For 7 days (from 7 days prior to time of occurrence)	Precipitation	Mean, sum

by combining and adjusting several variables from the raw data. If the variable name is not followed by a specific period, it is a 24 h weather statistic to predict frost.

Moreover, 70% of the randomly extracted data were used as training and validation data through cross-validation, and the remaining 30% of the data were evaluated as testing data. The ratio of frost occurrence and nonoccurrence in the training and testing data is presented in Table 3 and is classified as 1 : 3.

3.2. Model Fit Results. To avoid overfitting and underfitting in the training data, we used k-fold cross-validation. In k-fold cross-validation, the data were first partitioned into k equal (or nearly equal) segments or fold. Subsequently, k iterations of training and validation were performed such that, within each iteration, a different fold of the data was used for validation, whereas the remaining $k-1$-fold was used for learning [21]. As depicted in Figure 1, data were divided fivefold, calculating the average accuracy and area under the curve (AUC).

The average accuracy and AUC were selected as hyperparameters, and the model performance of the optimized hyperparameters was comprehensively evaluated based on various model evaluation criteria, such as the precision, recall, F1-score, CSI, accuracy, and AUC. The predicted and actual observations are expressed in a confusion matrix in Table 4.

Table 3: Count of frost occurrence and nonoccurrence in the dataset.

Dataset	Occurrence status	Count
Training and validation set	Frost (O)	21769
	Frost (X)	71387
Total training and validation set		93156
Testing set	Frost (O)	9290
	Frost (X)	30635
Total training and validation set		39925

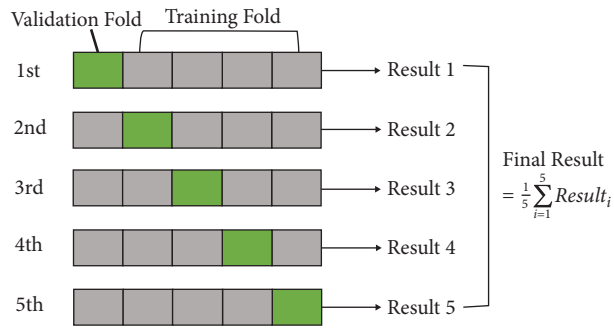

Figure 1: K-fold cross-validation process.

Table 4: Confusion matrix.

		Predicted	
		Frost (O)	Frost (X)
Actual	Frost (O)	True positive (TP)	False negative (FN)
	Frost (X)	False positive (FP)	True negative (TN)

The AUC refers to the area under the receiver operating characteristic (ROC) graph curve, revealing the performance of the classification model at all cut-off points. Its value was between 0.5 and 1, and a value closer to 1 indicates better model performance [22]. Accuracy indicates how well the model fits in the whole case. Precision is the rate of actual frost from predicted frost, and the recall value is the rate at which frost is predicted when actual frost is observed. The F1-score is a harmonized average of the precision and recall, primarily used when the data between classifications are severely unbalanced. The CSI is the hit rate of frost occurrence classifications excluding TNs. In natural conditions, far fewer frost phenomena cases occur than nonoccurrence cases. The CSI is considered the most important indicator because predicting frost occurrence is more important than predicting nonoccurrence [23]. Performance indicators closer to 1 indicate better performance. The evaluation indicators are defined in Equation (5) (except the AUC):

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN},\\
\text{Precision} &= \frac{TP}{TP + FP},\\
\text{Recall} &= \frac{TP}{TP + FN},\\
F1 &= \frac{2 \text{ times Precision times Recall}}{\text{Precision + Recall}},\\
\text{and CSI} &= \frac{TP}{TP + FP + FN}.
\end{aligned} \tag{5}$$

Table 5 lists the results verified with various hyperparameters. Although the performance was similar for each hyperparameter, the model adopted a maximum depth of 7 and a learning rate of 0.1, with the highest accuracy of 0.901 and an AUC of 0.944. The variable importance in the XGBoost model is depicted in Figure 2. The lowest temperature was the most important variable in XGBoost, followed by the minimum wind speed, standard deviation in temperature, maximum humidity, average wind speed, average precipitation, average cloudiness, average humidity, sum of total solar radiation time, maximum cloudiness, and cumulative precipitation over 3 days.

3.3. Comparison of Model Performance Evaluation Using K-Means Clustering. The XGBoost model was applied to the testing data with the selected hyperparameters. When the model performance evaluation index was calculated for each station point, the performance greatly differed for each point. However, some observation points had fewer than 100 cases due to randomization, so the reliability of evaluating the model performance by branch was low. Thus, the model performance was evaluated by dividing groups using the *k*-means method with weather variable characteristics.

In addition, *k*-means clustering was calculated with the average value of the 24 h minimum temperature, minimum wind speed, standard deviation in temperature, maximum humidity, average wind speed, average precipitation, and average cloudiness, which were important variables in XGBoost. The calculation of the silhouette score to determine the number of clusters is presented in Figure 3, and accordingly, clustering was classified into two groups by the highest silhouette value.

The characteristics of the classified group were drawn in a boxplot, as illustrated in Figure 4. Group 1 had a high minimum temperature, and the standard deviation in temperature was lower than that of Group 2. The average precipitation, clouds, humidity, and wind speed-related indices were higher than those for Group 2, which seems to be close to the coastal climate. The visualization in Figure 5(a) indicates that, by reducing the data dimension using the principal component analysis (PCA), the island belongs only to Group 1, and 75% of Group 1 is an island. Group 1 was judged to be a coastal climate, and the distance from each observation point to the nearest coast and the altitude of the observation point were calculated (Figure 5(b)). For Group 1, the observation point and coastline were close for many points, and the altitude was low.

The frequency of frost occurrence (number of frost occurrences/total cases) was also about 6.3% in Group 1 and 28.9% in Group 2, which was much higher. It was confirmed that the distribution pattern of weather factors varies depending on the location, and the frost incidence rate varies accordingly. Therefore, the points with strong coastal climate characteristics were classified into Group 1, and those with inland climate characteristics were classified into Group 2. The accuracy was calculated by classifying them as existing results, as presented in Table 6.

The accuracy results were very different depending on the group. For Group 1, the accuracy was high, but the precision, recall, F1-score, and CSI scores were lower than those of Group 2. In particular, the CSI was 0.346 in Group 1 and 0.668 in Group 2, which was about twice the difference.

3.4. Comparison of Model Performance considering the Frost Occurrence Environment. The *k*-means clustering method confirmed that the distribution pattern of the meteorological factors was very different depending on the geographical

Table 5: Optimization hyperparameter validation.

Max depth	Learning rate	AUC (area under the Curve)	Accuracy
3	0.1	0.938	0.895
7	**0.1**	**0.944**	**0.901**
10	0.1	0.944	0.901
3	0.5	0.942	0.900
7	0.5	0.938	0.897
10	0.5	0.936	0.896
3	1	0.938	0.896
7	1	0.924	0.883
10	1	0.926	0.887

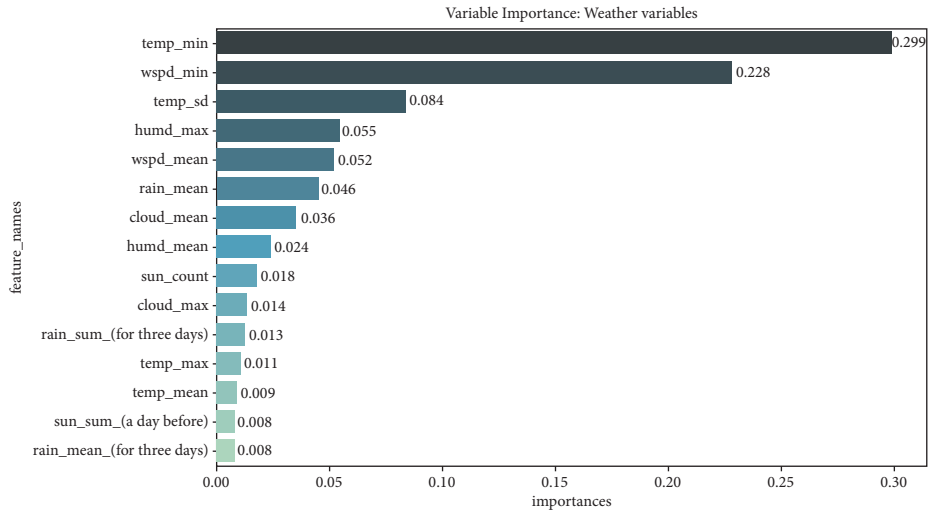

Figure 2: Variable importance of weather variables in the XGBoost model.

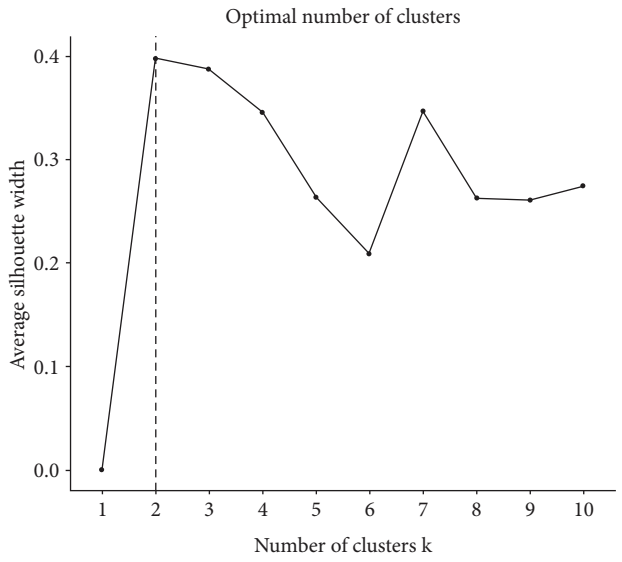

Figure 3: Calculation of the silhouette score in k-means clustering.

characteristics. The distance from the observation point to the nearest shoreline and the altitude of the observation point were reflected as additional input variables to reflect the geographic characteristics modeled on XGBoost. The top 15 variables of importance are provided in Figure 6(a). The distance to the coastline and altitude were selected as the 9th and 11th important variables, respectively, and wind speed (min/average), temperature (min/average), humidity (maximum/average), precipitation average, average cloudiness, and total solar radiation time were still important variables.

The model result is presented in Table 7. Accuracy rose slightly compared to when only the weather variable was considered, but the values of precision in Group 1 and the precision and recall in Group 2 increased, and as a result, the F1-score and CSI values in both groups increased. Overall, geographic characteristics are valid variables in predicting frost, considering that the values of all indicators improved.

In addition, considering that the frequency of frost occurrence greatly differed depending on the group, this study proposes a method to determine this information. First, once frost occurs, frost is likely to occur the next day; therefore, a linear correlation was examined to determine the continuity of frost occurrence. The linear correlation coefficient was quite high at 0.65, with the previous day's frost occurrence variable and the day's frost occurrence variable. The previous day's frost occurrence was reflected as an input variable, and the model result is presented in Table 8. All indicators including the F1-score and CSI values of Group 1 and Group 2 increased significantly. The variable

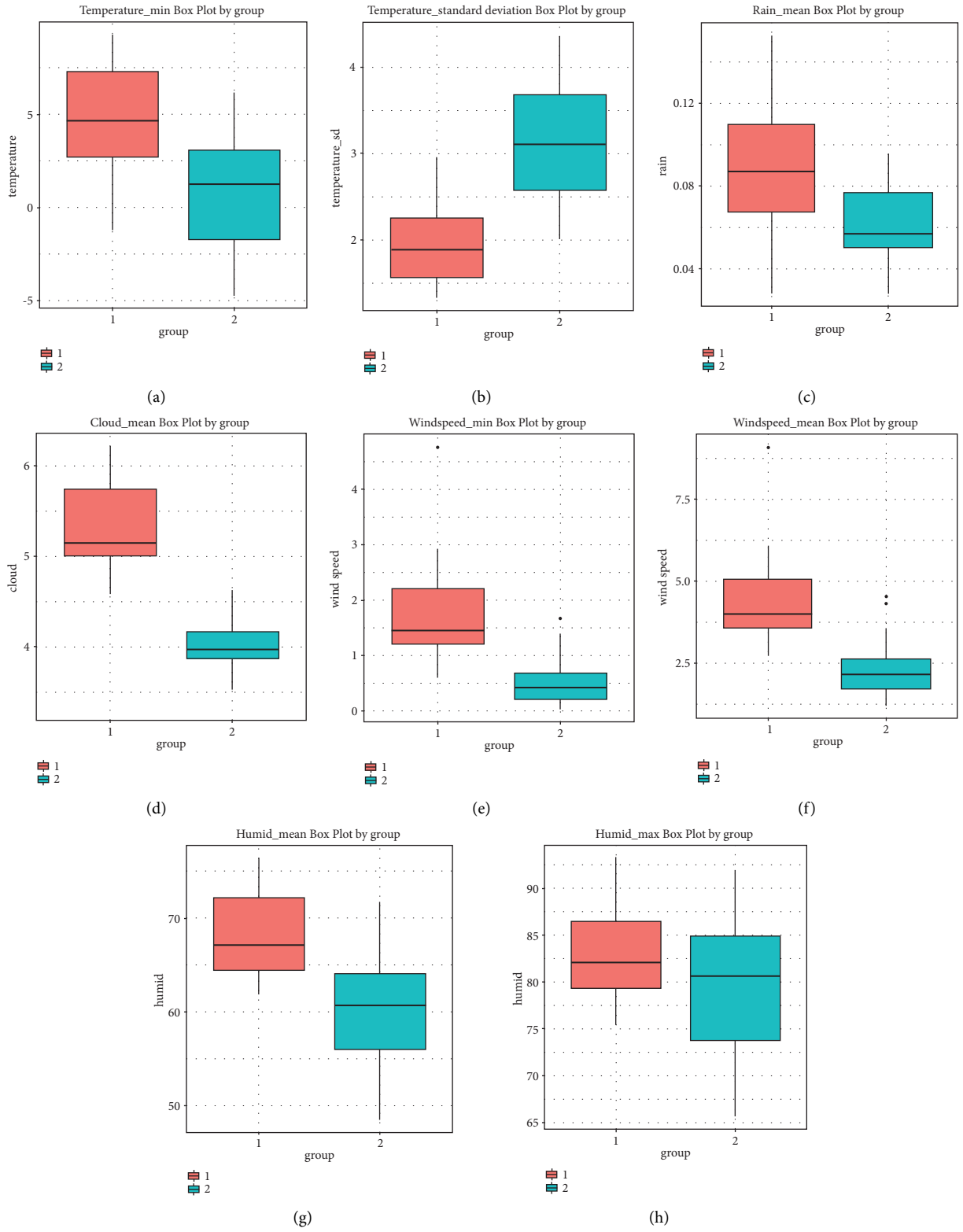

Figure 4: Box plot by group: (a) minimum temperature, (b) standard deviation in temperature, (c) average precipitation, (d) average cloudiness, (e) minimum wind speed, (f) average wind speed, (g) average humidity, and (h) maximum humidity.

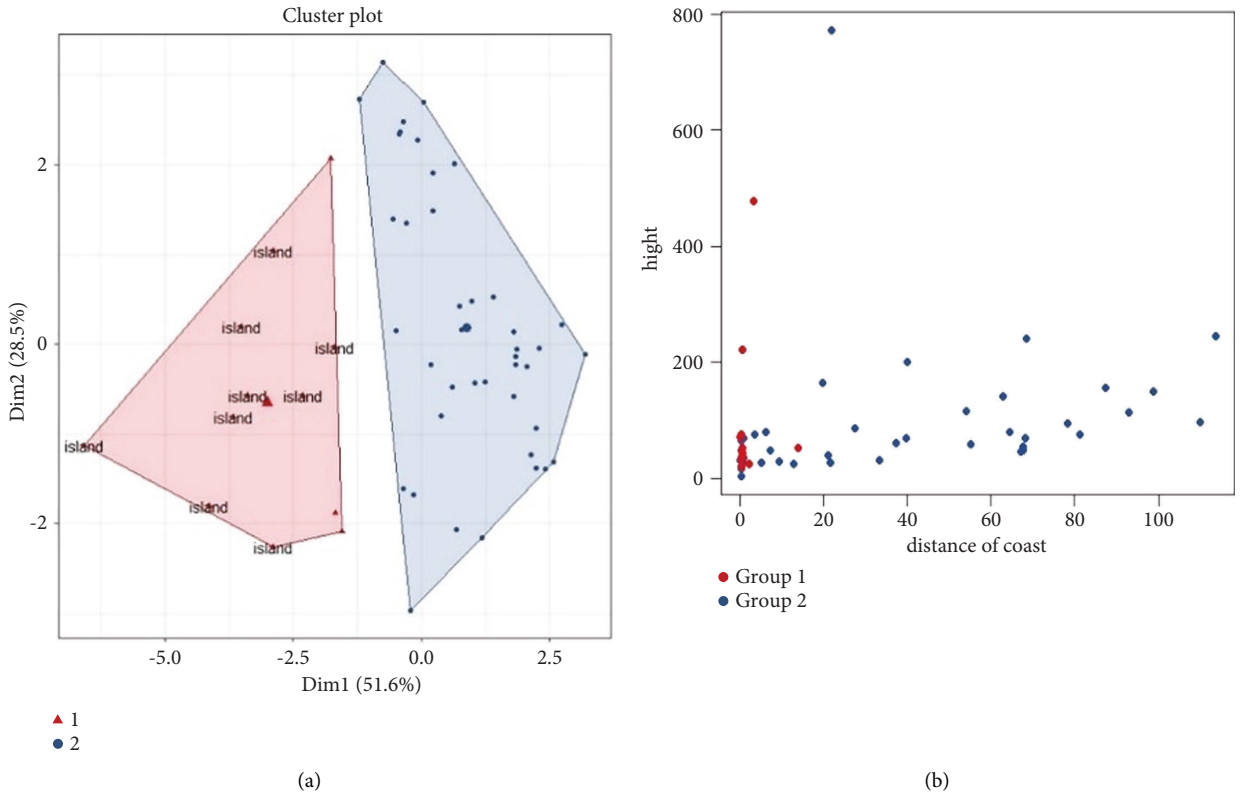

Figure 5: (a) *K*-means clustering visualization using the principal component analysis and (b) coastal distance and altitude scatter plot by group.

Table 6: Results by group with XGBoost using weather variables.

Group	TN (true negative)	FP (false positive)	FN (false negative)	TP (true positive)	Accuracy	Precision	Recall	F1	CSI
1	9174	144	358	266	0.950	0.649	0.426	0.515	0.346
2	19982	1335	1987	6679	0.889	0.833	0.771	0.801	0.668
Total	29156	1479	2345	6945	0.904	0.824	0.748	0.784	0.645

importance (top 15 variables) of the XGBoost model with frost occurrence the previous day added as a variable is presented in Figure 6(b). Frost from the previous day was overwhelmingly considered the most important variable, followed by the temperature (standard deviation), lowest wind speed, and maximum humidity.

Second, the frequency of frost differs greatly between coastal and inland areas; thus, it was decided to check whether learning was particularly poor for the coastal areas. The imbalance of the frost generation case was oversampled with SMOTE to create balanced data. In addition, SMOTE provides more related minority class samples to learn from, allowing a learner to carve broader decision regions, leading to more coverage of the minority class [24]. In the training data, the frost rate in coastal areas was 6.3%, and in inland areas, the rate was 28.9%; therefore, it would be challenging to generate frost cases in coastal areas to oversample using SMOTE. Thus, coastal and inland areas were separated from the training data and preprocessed using SMOTE, correcting the imbalance in frost data.

In addition, SMOTE increases the rate of frost detection (recall). The number of frost occurrences in the training data was adjusted to be the same as the number of nonfrost occurrences. Then, the coastal and inland areas were modeled separately with XGBoost using weather and geographic factors and the previous day's frost as input variables. The top 15 important variables are presented in Figure 7. The lowest temperature and whether frost occurred the previous day are still important variables for both groups. However, there was a ranking difference in the variable importance between coastal and inland regions. With coastal characteristics, Group 1 demonstrated a higher importance of geographic variables (distance from coastline and altitude) and variables to recognize land conditions, such as past precipitation information (total precipitation the previous day, total for 3 days, and total for 7 days).

The result of the model fit is in Table 9. The recall scores rose for coastal and inland areas, but the accuracy, F1-score, and CSI score fell due to a significant decrease in precision. Increasing the frost generation case with SMOTE made it possible to increase the number of cases predicting frost, but the increase in FP affected other performance indicators. As a result, when comprehensively examining the F1-score and

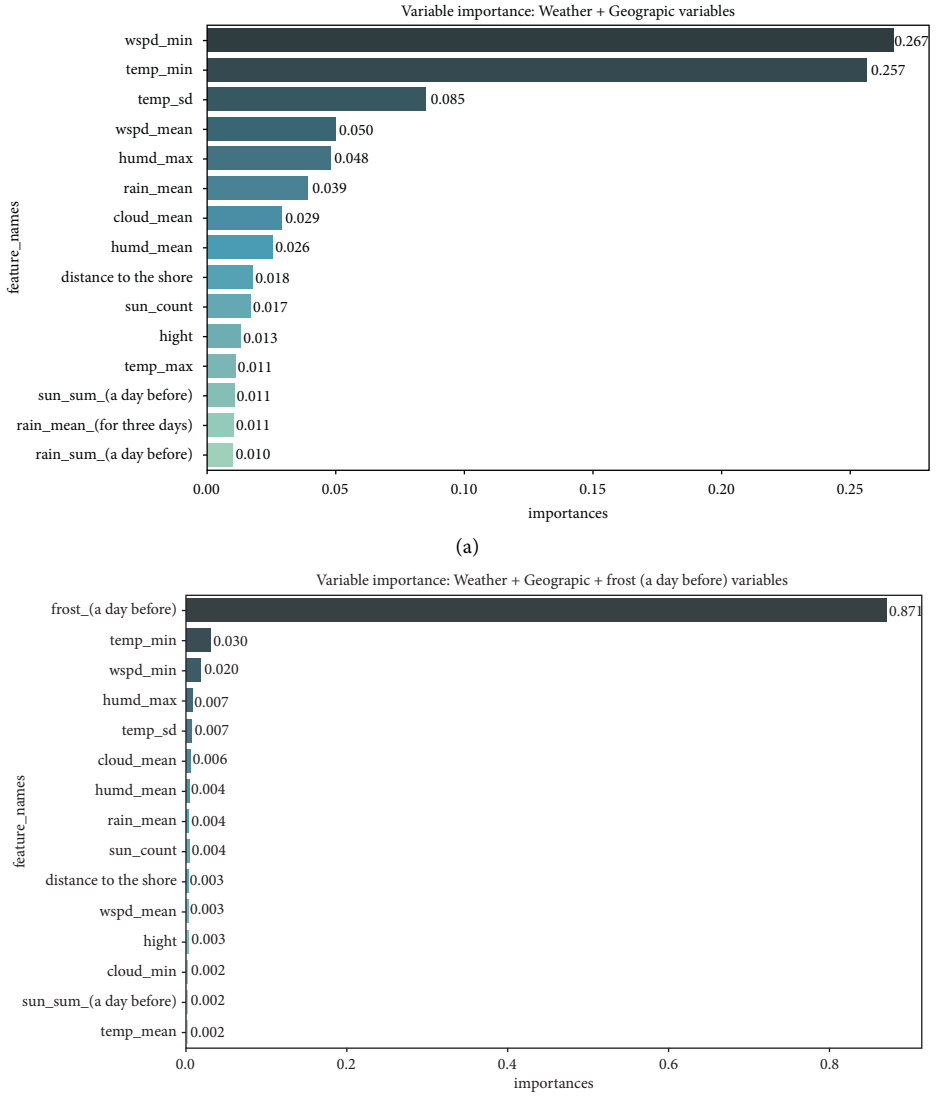

Figure 6: (a) Variable importance of weather and geographic variables in the XGBoost model and (b) variable importance of weather, geographic, and frost (a day before) variables in the XGBoost model.

Table 7: Results by group with XGBoost using weather and geographic variables.

Group	TN (true negative)	FP (false positive)	FN (false negative)	TP (true positive)	Accuracy	Precision	Recall	F1	CSI
1	9205	113	364	260	0.952	0.697	0.417	0.522	0.353
2	19959	1358	1837	6829	0.893	0.834	0.788	0.810	0.681
Total	29164	1471	2201	7089	0.908	0.828	0.763	0.794	0.659

Table 8: Results with XGBoost by group using weather, geographic, and frost (a day before) variables.

Group	TN (true negative)	FP (false positive)	FN (false negative)	TP (true positive)	Accuracy	Precision	Recall	F1	CSI
1	9204	114	296	328	0.959	0.742	0.526	0.615	0.444
2	20320	997	1562	7104	0.915	0.877	0.820	0.847	0.735
Total	29524	1111	1858	7432	0.926	0.870	0.800	0.834	0.715

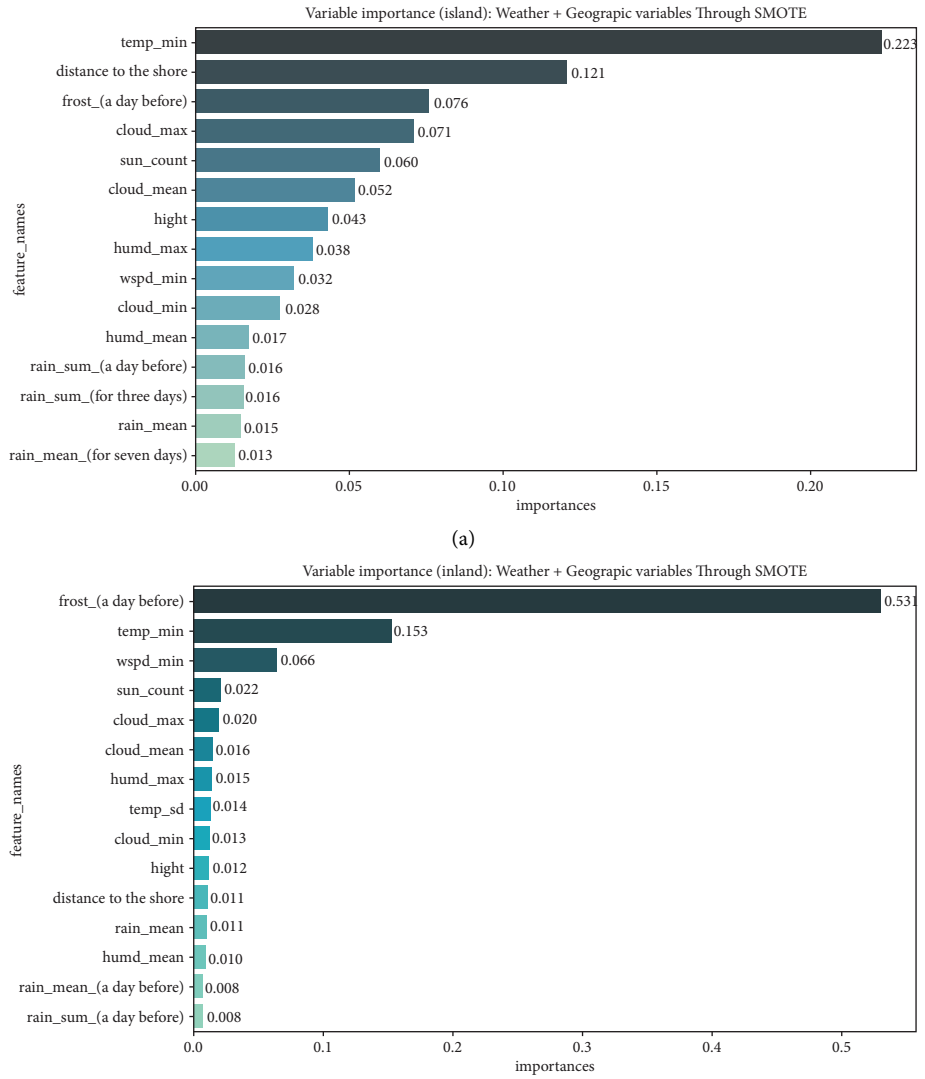

Figure 7: (a) Variable importance of XGBoost with all variables using SMOTE techniques in Group 1 and (b) variable importance of XGBoost with all variables using SMOTE techniques in Group 2.

Table 9: Results of XGBoost by group with all variables using the SMOTE technique.

Group	TN	FP	FN	TP	Accuracy	Precision	Recall	F1	CSI
1	8995	323	225	399	0.945	0.553	0.639	0.593	0.421
2	19953	1334	1402	7264	0.909	0.845	0.838	0.842	0.726
Total	28948	1657	1627	7663	0.918	0.822	0.825	0.824	0.700

Table 10: Summary of XGBoost results by group.

Group		SMOTE X			SMOTE O
		Weather	Weather + geographic	Weather + geographic + previous day's frost	Weather + geographic + previous day's frost
Island	Accuracy	0.950	0.952	0.959	0.945
	Precision	0.649	0.697	0.742	0.553
	Recall	0.426	0.417	0.526	0.639
	F1-score	0.515	0.522	0.615	0.593
	CSI	0.346	0.353	0.444	0.421
Inland	Accuracy	0.889	0.893	0.915	0.909
	Precision	0.833	0.834	0.877	0.845
	Recall	0.771	0.788	0.820	0.838
	F1-score	0.801	0.810	0.847	0.842
	CSI	0.668	0.681	0.735	0.726

the CSI index, which is sensitive to both FP and FN, SMOTE was used to increase the learning rate, but the performance was not significantly improved.

4. Conclusion

As displayed in Table 10, summarizing the frost prediction results of XGBoost, the regional accuracy was significantly different. With weather variables, the observation stations were divided into two groups to determine the characteristics of the group. Group 1 included 75% of the islands. In addition, Group 1 had a coastal climate, considering that the standard deviation in temperature was lower than that for Group 2, and the lowest temperature, average precipitation, cloud volume, humidity, and wind speed-related indices were higher. As a result of separately calculating the accuracy of regions with coastal climatic characteristics and regions with inland climatic characteristics, it was found that the precision and recall scores of coastal regions were much lower than those of inland regions.

First, geographic elements were added to the weather variables to improve accuracy. For the F1-score and CSI, the comprehensive indices of precision and recall, the coastal and inland areas improved.

Next, the difference in the frost occurrence rate was supplemented. To this end, the continuity of the frost generation was considered. Whether frost occurred the previous day was added as an input variable, and as a result, the scores were greatly improved in all performance indicators for coastal and inland areas. In particular, the CSI increased by about 10% in coastal areas, from 35.3% to 44.4%. Second, the SMOTE technique was used to improve learning ability. Although more predictions of frost were made, the precision score dropped significantly as the prediction was often wrong. However, by fitting the models by group in the SMOTE approach, we found that there is a significant ranking difference of variable importance depending on the group. Through this, it is necessary to understand and apply the geographical characteristics in predicting frost occurrence.

Based on the results of various studies, it was confirmed that the characteristics of the terrain affect the occurrence of frost, and the contribution of frost occurrence differs for each terrain. If specific standards for important factors for frost occurrence are established by region, accuracy can be improved through the reorganization of variables. In the future, various approaches to improving frost prediction are expected to contribute to efficient frost alarms and preventive activities.

Abbreviations

AUC:	Area under the curve
CSI:	Critical success index
FN:	False negative
FP:	False positive
MLP:	Multilayer perceptron
PCA:	Principal component analysis
RF:	Random forest
ROC:	Receiver operating characteristic
SVM:	Support vector machine
SMOTE:	Synthetic minority oversampling technique
TN:	True negative
TP:	True positive
XGBoost:	EXtreme gradient boosting (only used in variable importance plot)
Temp:	Temperature
Wspd:	Wind speed
Humd or Humid:	Humidity
Cloud:	Cloudiness
Sun:	Solar radiation
Rain:	Precipitation
Min:	Minimum
Max:	Maximum.

Data Availability

The weather observation data used to support the findings of this study may be released upon application to the Korea Meteorological Administration, who can be contacted at https://data.kma.go.kr.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B01014954). This research was supported by the Chung-Ang University Research Grants in 2021.

References

[1] Y. B. Lee and S. T. Ro, "Frost formation on a vertical plate in simultaneously developing flow," *Experimental Thermal and Fluid Science*, vol. 26, no. 8, pp. 939–945, 2002.

[2] F. M. DaMatta and J. D. C. Ramalho, "Impacts of drought and temperature stress on coffee physiology and production: a review," *Brazilian Journal of Plant Physiology*, vol. 18, no. 1, pp. 55–81, 2006.

[3] J. R. Lamichhane, "Rising risks of late-spring frosts in a changing climate," *Nature Climate Change*, vol. 11, no. 7, pp. 554-555, 2021.

[4] S. J. Crimp, B. Zheng, N. Khimashia et al., "Recent changes in southern Australian frost occurrence: implications for wheat production risk," *Crop & Pasture Science*, vol. 67, no. 8, pp. 801–811, 2016.

[5] L. Ghielmi and E. Eccel, "Descriptive models and artificial neural networks for spring frost prediction in an agricultural mountain area," *Computers and Electronics in Agriculture*, vol. 54, no. 2, pp. 101–114, 2006.

[6] P. Sallis, M. Jarur, and M. Trujillo, "Frost prediction characteristics and classification using computational neural networks," in *International Conference on Neural Information Processing*, pp. 1211–1220, Springer, Berlin, Germany, 2008, November.

[7] H. Lee, J. A. Chun, H. H. Han, and S. Kim, "Prediction of frost occurrences using statistical modeling approaches," *Advances in Meteorology*, vol. 2016, Article ID 2075186, 9 pages, 2016.

[8] A. Castañeda-Miranda and V. M. Castaño, "Smart frost control in greenhouses by neural networks models," *Computers and Electronics in Agriculture*, vol. 137, pp. 102–114, 2017.

[9] É. S. Diniz, A. S. Lorenzon, N. L. M. de Castro et al., "Forecasting frost risk in forest plantations by the combination of spatial data and machine learning algorithms," *Agricultural and Forest Meteorology*, vol. 306, Article ID 108450, 2021.

[10] A. Zendehboudi and X. Li, "Robust predictive models for estimating frost deposition on horizontal and parallel surfaces," *International Journal of Refrigeration*, vol. 80, pp. 225–237, 2017.

[11] A. L. Diedrichs, F. Bromberg, D. Dujovne, K. Brun-Laguna, and T. Watteyne, "Prediction of frost events using machine learning and IoT sensing devices," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4589–4597, 2018.

[12] M. Rostamian and A. H. Halabian, "Statistical analysis of the frequency and stability of the frost days in southern Khorasan Province, using Markov chain Model," *Spatial Planning*, vol. 8, no. 2, pp. 39–60, 2018.

[13] L. Ding, K. Noborio, and K. Shibuya, "Frost forecast using machine learning-from association to causality," *Procedia Computer Science*, vol. 159, pp. 1001–1010, 2019.

[14] Y. Tamura, L. Ding, K. Noborio, and K. Shibuya, "Frost prediction for vineyard using machine learning," in *Proceedings of the 2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium On Advanced Intelligent Systems (SCIS-ISIS)*, pp. 1–4, Hachijo Island, Japan, 2020.

[15] J. R. Rozante, E. R. Gutierrez, P. L. Silva Dias, A. Almeida Fernandes, D. S. Alvim, and V. M. Silva, "Development of an index for frost prediction: technique and validation," *Meteorological Applications*, vol. 27, no. 1, Article ID e1807, 2020.

[16] S. Wassan, C. Xi, N. Z. Jhanjhi, and L. Binte-Imran, "Effect of frost on plants, leaves, and forecast of frost events using convolutional neural networks," *International Journal of Distributed Sensor Networks*, vol. 17, no. 10, Article ID 155014772110537, 2021.

[17] H. Kim and S. Kim, "A study on frost prediction model using machine learning," *The Korean journal of applied statistics*, vol. 35, no. 4, pp. 543–552, 2022.

[18] H. Zheng and Y. Wu, "A XGBoost model with weather similarity analysis and feature engineering for short-term wind power forecasting," *Applied Sciences*, vol. 9, no. 15, p. 3019, 2019.

[19] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768-769, 1965.

[20] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747-748, Sydney, NSW, Australia, 2020.

[21] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[22] M. P. Muller, G. Tomlinson, T. J. Marrie et al., "Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia?" *Clinical Infectious Diseases*, vol. 40, no. 8, pp. 1079–1086, 2005.

[23] I. Noh, H. W. Doh, S. O. Kim, S. H. Kim, S. Shin, and S. J. Lee, "Machine learning-based hourly frost-prediction system optimized for orchards using automatic weather station and digital camera image data," *Atmosphere*, vol. 12, no. 7, p. 846, 2021.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.