

Assessment of Machine Learning Algorithms for Land Cover Classification Using Remotely Sensed Data

Jeongmook Park,^{1†} Yongkyu Lee,^{2†} and Jungsoo Lee^{2*}

¹Researcher, Forest ICT Research Center, National Institute of Forest Science

²Department of Forest Management, Kangwon National University, Chuncheon 24341, Republic of Korea

(Received September 1, 2021; accepted November 9, 2021)

Keywords: machine learning, optimization, land cover map, random forest, XGBoost, LightGBM

The purpose of this study was to apply the random forest (RF), XGBoost, and LightGBM machine learning (ML) algorithms to land cover classification, and to present the model tuning process for each algorithm. Sentinel-2 satellite images were used for land cover classification, and the land cover map provided by the Ministry of Environment of the Republic of Korea was used as label data. Each ML algorithm was applied using the constructed dataset. In addition, each ML algorithm was optimized by three methods (grid search, random search, and Bayesian optimization). The grid search took the longest time to optimize the hyperparameters because it required the highest number of search iterations, but the accuracy was highest. The random search was the fastest method of optimizing the hyperparameters. The accuracy of XGBoost was the highest for each ML algorithm. The prediction of XGBoost was the most consistent with the land cover map provided by the Ministry of Environment. However, the LightGBM algorithm has a major advantage in terms of the algorithm optimization and application time. Therefore, our study is meaningful in that we obtained a higher accuracy and shorter time for each ML algorithm.

1. Introduction

With the annual changes and updates of the greenhouse gas (GHG) inventory in the land use, land use change, and forestry (LULUCF) categories, the systematic monitoring of the area maintained in each category is required along with the monitoring of its changes.⁽¹⁾ For this purpose, the application of a geographic information system and remote sensing (GIS and RS) using satellite imagery and spatial data is instrumental. In particular, the GHG inventory development through the application of GIS/RS has merit in terms of accurate measurement of uncertainties, as it enables the development of the inventory at the Approach 3 level in the IPCC Guidelines.⁽²⁾ Accordingly, research on changes in land use and land cover using satellite imagery has been continuously ongoing in accordance with the increasing public interest in artificial intelligence (AI) since the late 2010s. In addition, the recent interest in the application of machine learning (ML) technology has also increased in the field of remote sensing.⁽³⁾

*Corresponding author: e-mail: jslee72@kangwon.ac.kr

†These authors contributed equally to this work.

<https://doi.org/10.18494/SAM.2021.3612>

Presently, ML is one of the fastest growing fields of technology and has a multidisciplinary approach owing to the adoption of computer science and statistics. ML technology can be viewed as playing a pivotal role in AI and data science. ML technology is now employed for decision-making in various fields beyond science and technology.^(4,5)

Since the launch of the Landsat satellite in 1972, various ML algorithms have been utilized for pixel-based image classification, such as the maximum likelihood method, ISODATA, K-means, support vector machine (SVM), and artificial neural networks. In particular, active research has been undertaken to apply the random forest (RF) algorithm for land cover classification.^(6,7) In addition, it has been reported that the RF algorithm has a higher potential than existing land cover classification algorithms such as the maximum likelihood method, minimum distance method, decision trees, and SVM. In addition, land cover classification studies using the XGBoost and LightGBM algorithms based on decision trees, such as RF, have been conducted since 2018. In particular, a limited number of studies have been undertaken on land cover classification using the LightGBM algorithm.^(8–10) For the above three algorithms, high accuracy can be achieved only through hyperparameter (HP) optimization for the dataset.⁽¹¹⁾ However, in land cover classification studies, the process of model tuning is omitted or described very briefly, leading to a lack of information about the HP optimization process in the field of remote sensing.

Therefore, in this study, a land cover map was constructed by applying the RF, XGBoost, and LightGBM algorithms, which are widely used in the field of remote sensing. In addition, a model tuning process was used for each algorithm so that the optimization method could be selected depending on the purpose.

2. Materials

2.1 Sentinel-2 satellite imagery

Satellite imagery from Sentinel-2 was utilized for the present study. The swath of the Sentinel-2 satellite was 290 km, the radiometric resolution was 12 bits, and the temporal resolution was 5 days (based on the equator). The images consisted of 13 bands (four 10 m bands, six 20 m bands, and three 60 m bands). Sentinel-2 satellite imagery provides more information in near-infrared (NIR) and shortwave infrared (SWIR) bands than Landsat 8 satellite imagery. Therefore, it is available for monitoring urban areas with precision level. In addition, land cover classification can be performed owing to its high temporal resolution.⁽¹²⁾ Sentinel-2 imagery complements Landsat satellite imagery and SPOT satellite imagery to increase data availability.^(13,14)

2.2 Land cover map by Ministry of Environment

The Ministry of Environment of the Republic of Korea has constructed a land cover map for the entire area of the national territory since 1998, starting with the level-1 land cover map and further developing and distributing the level-2 and level-3 land cover maps. The maps are

classified into 7, 22, and 41 categories for the level-1, level-2, and level-3 classifications, respectively. The level-1, level-2, and level-3 classifications have spatial resolutions of 30, 5, and 1 m, respectively. The level-2 land cover map, which has the same spatial resolution as the Sentinel-2 imagery, has been utilized to convert level-2 classification items on the level-3 land cover map since the last update of the map in 2013. Therefore, to match the time series with the Sentinel-2 imagery implemented in this study, the level-3 land cover map was utilized. As the categories, the following categories of the level-1 land cover map were adopted: used area, agricultural land, forest, grass, wetland, barren, and water.⁽¹⁵⁾

2.3 Study area

For the study area, Wonju-si was selected, which is located on the administrative border of Gangwon-do, Gyeonggi-do, and Chungcheongbuk-do in the Republic of Korea (Fig. 1, 127° 44–128° 12 E, 37° 08–37° 31 N). The total area of Wonju-si is approximately 86,644 ha. Among the 18 cities and counties in Gangwon-do, not only has the ratio of forest area in Wonju-si shown the most dramatic decrease but also the population has increased by about 6% over the past 10 years (Fig. 1).

3. Methodology

The data implemented in the present study were constructed through the preprocessing of Sentinel-2 image data. In addition, spectral values, gray level co-occurrence matrix (GLCM) values, and topographic data were constructed. The GLCM values were derived from the

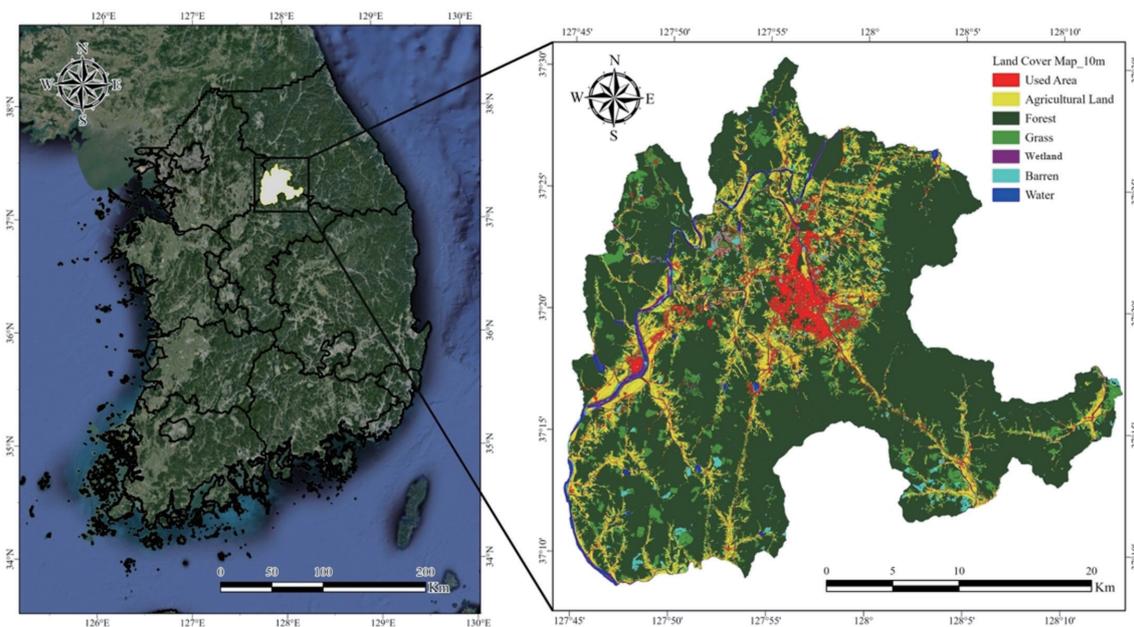


Fig. 1. (Color online) Study area (Wonju-si, land cover map provided by Ministry of Environment).

spectral values by principal component analysis (PCA). The label data for training were constructed by implementing the land cover map of the Ministry of Environment. The land cover classification algorithms applied in this study were RF, XGBoost, and LightGBM, which are ML-based algorithms. In addition, the optimal classification algorithm was selected after optimization. The classification model and model tuning were applied using the ML Python scikit-learn library. Furthermore, a land cover map of Wonju-si was constructed using the optimal model, and its consistency with the land cover map of the Ministry of Environment of Wonju-si was examined (Fig. 2).

3.1 Preprocessing of remote sensing data

Using Sentinel-2 imagery, class differentiation can be maximized owing to the large seasonal variation in areas with vegetation. Images taken on January 3, 2019 and September 30, 2019 were utilized to analyze the characteristics of the growth and the non-growth period.⁽¹⁶⁾ The images of the study area on the day the images were captured were easy to analyze owing to the absence of clouds. The correction level of Sentinel-2 imagery was Level-1C. Sentinel-2 Level-1C is an ortho-image product composed of 100×100 km² tiles of the UTM/WGS84 projection. In addition, for geometric correction, a 90 m digital elevation model (DEM) was utilized for resampling. Therefore, the spectral values were directly utilized for analysis without any pretreatment. Next, for a consistent analysis, resampling was performed at a spatial resolution of 10 m for bands with a spatial resolution of 20 m. In addition, to cover the study area, four satellite images were mosaiced, and extraction by a mask was performed in the study area.

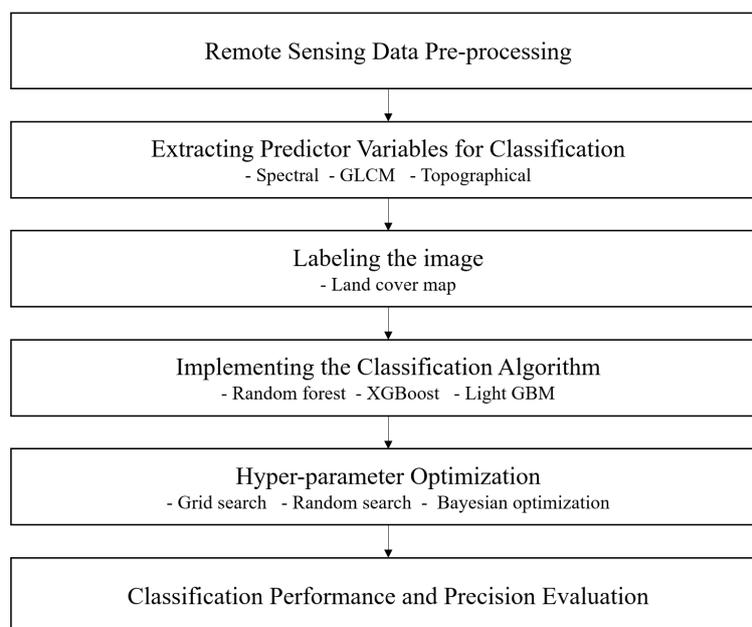


Fig. 2. Flowchart of overall research procedure.

3.2 Selection of input variables for ML algorithms

The spectral values of satellite imagery were utilized as input variables in the study of land use and cover classification using the ML algorithms. In this study, the GLCM variable, texture data, and topographic variables were implemented along with spectral values, which are commonly utilized in land use and land cover classification. The GLCM variable was derived by the PCA of the spectral values. The GLCM can improve the classification accuracy of the used area and wetland. This method was applied on the basis of previous reports indicating that an appropriate result for the classification of land cover is provided when the first principal component is utilized.^(17,18) Topographical variables, including the slope and aspect derived from the DEM, were used as input variables in the classification process, primarily to reduce the misclassification of shadowed areas.⁽¹⁹⁾ Therefore, a total of 39 input variables were utilized in this study: ten spectral variables of the Sentinel-2 imagery and eight GLCM variables of PCA1 for both the growth and non-growth periods, and the three topographical variables constructed with the aid of the DEM.

3.3 Labeling the image

The dataset is an important component in satellite-image-based land cover classification. To construct the dataset for the application of each ML algorithm, 86644 (approximately 1%) of the 8664400 pixels were randomly sampled, corresponding to approximately 1% of the total study area.⁽²⁰⁾ The dataset extracted through this process was randomly divided in a ratio of 7:3 for training and validation, and the testing of each ML algorithm. The constructed dataset is presented in Table 1.

3.4 Implementing the ML algorithm for land cover classification

For the land cover classification algorithm, ensemble algorithms based on the decision tree were utilized with the ML algorithms. A decision tree is a non-parametric method that does not require assumptions, such as the linearity, normality, and equal variance of the data, and is highly beneficial for general types of data.⁽²¹⁾ Ensemble algorithms based on decision trees are mainly divided into bagging and boosting algorithms. In this study, the RF algorithm (based on

Table 1
Land cover classes and dataset pixel numbers collected for the classification.

Class	Dataset	
	Training and validation	Testing
Used area	3412	1455
Agricultural land	6758	2977
Forest	40553	17419
Grass	6449	2679
Wetland	888	365
Barren	1943	823
Water	647	276

bagging) and the XGBoost and LightGBM algorithms (based on boosting) were utilized. RF has been extensively applied to the field of remote sensing since the 2000s and has been adopted for land cover classification.^(6,22) The high applicability of RF to land cover classification has already been demonstrated.⁽²³⁾ XGBoost and LightGBM have relatively been developed recently, are based on the gradient boosting decision tree, and have the advantages of short training time and high accuracy. These three algorithms were utilized owing to their strengths, and the results were compared.⁽²⁴⁾

RF is a bagging-type ensemble algorithm that synthesizes predictions utilizing multiple decision trees. The RF classifier generally has a higher classification accuracy than a single decision tree. In particular, RF is effective when using small sample sizes with high-dimensional data input.⁽²⁵⁾ RF is the integrated form of the bootstrap, random subspace, and aggregation algorithms. In particular, the bootstrap and aggregation algorithms are also called bagging algorithms.^(26,27)

XGBoost is a decision-tree-based ensemble ML algorithm and is a popular gradient boosting library with parallel computation. In bagging, each decision tree is connected in parallel to derive different data and results. However, in gradient boosting, each decision tree is connected hierarchically and the result obtained from the input to the individual decision trees is reinstated to other individual decision trees. That is, a new predictor is trained on the basis of the residual error derived from the previous individual decision tree. XGBoost provides an environment that enables the parallel processing of serial-structured algorithms.^(28,29)

LightGBM is an algorithm with a learning time that is greatly reduced by implementing a gradient boosting algorithm with a lighter computational load than other ensemble algorithms, and its memory usage is low. A characteristic of LightGBM is that, whereas existing decision tree algorithms equally split the depth of trees, LightGBM does not consider the depth of trees but splits the node with the maximum value of loss to build an asymmetric tree. Therefore, the time required to equally divide the tree depth is reduced, but the risk of overfitting is greater than that of other decision-tree-based algorithms. Therefore, the application of the LightGBM algorithm is not recommended for less than 10000 items of data because of the high risk of overfitting.⁽³⁰⁾

3.5 HP optimization with grid search, random search, and Bayesian optimization

A parameter refers to a variable used to determine the functional relationship between several variables. In ML, parameters indicate values determined internally by the model, and HPs indicate values directly set by a researcher in modeling. In general, as there is no set way for finding an optimal value for an HP, it is determined by a subjective empirical method, termed as a manual search. The drawbacks of this method are that there is no guarantee that the value of an HP determined empirically is the optimal value in the actual model and that determining the optimal value becomes more difficult with increasing number of HPs. In addition, many ML models in the field of remote sensing in previous studies were fit using default parameters. However, there has not yet been a thorough examination of the parameter sensitivity. We analyzed the relationship between parameters and performance.^(10,24,31) To address these

limitations, we applied all three methods of grid search, random search, and Bayesian optimization, which are widely used as HP optimization methods. The optimal HP selected for each method, the search iteration, the required time, and the accuracy are presented, and each optimization method was comparatively analyzed to enable its appropriate selection and utilization in future research.

In a grid search, the researcher sets up a grid of HP values and builds a model for every combination of HPs for a comparative accuracy assessment. The grid search method has the advantage of being able to perform an exhaustive search for all values that HPs can take. However, it also has a disadvantage in that the required computation time increases exponentially with the number of HPs and the optimal values may never be revealed depending on the interval between HP values.

In a random search, the range of values is specified for each HP, and the values of the HPs are randomly selected and compared. A random search has an advantage in that the number of search iterations can be directly specified to adjust the required time for computation, and the optimal parameter within the HP range can also be searched for. On the other hand, the method has a disadvantage of not utilizing the previous data because the values are randomly selected for every search.

Bayesian optimization is a model optimization technique that defines an objective function and determines the optimal solution of the HP to maximize the objective function. Bayesian optimization allows the systematic search of HPs utilizing previously searched information.

In the HP optimization process, the two HPs tuned in the RF algorithm are `n_estimators` and `min_samples_split`, where `n_estimators` is the number of trees in a forest, that is, the number of trees to be built, and `min_samples_split` is the minimum number of samples of the node at node splitting. The five HPs tuned in the XGBoost algorithm are `max_depth`, `min_child_weight`, `gamma`, `subsample`, and `learning_rate`: `max_depth` is the maximum depth when building a tree; `min_child_weight` is the number of samples used for additional partitioning in a tree and is utilized to prevent overfitting; `gamma` is also adopted in additional partitioning in a tree, and the larger the value, the less likely a node will be split in a tree, resulting in a conservative model; `subsample` is the ratio when creating a subset from the original dataset and is also implemented to prevent overfitting; `learning_rate` is the weight of the new tree in the gradient boosting model, and overfitting is prevented by reducing the learning rate in the gradient boosting model. For LightGBM, the same HPs as those in XGBoost were selected except for `gamma`, which does not correspond to an HP of LightGBM.

The search range was set on the basis of an existing study that performed HP optimization using a grid search.⁽³²⁾ For RF and XGBoost, the search range was set to be the same as in previous studies, and the search range for LightGBM was set to be the same as that for XGBoost. In addition, the `gamma` parameter was excluded because of its inapplicability. In addition, the search iteration for the random search and Bayesian optimization was set to 75 with reference to the analysis of previous studies using ML algorithms in other fields (Table 2).⁽³³⁾

K-fold cross-validation was adopted to select the optimum HP. This method can obtain 0.1–3% more accurate results than the hold-out method, which performs training and validation only once.⁽³⁴⁾ *K*-fold cross-validation divides the training and validation datasets into *K* subsets,

Table 2
Searches for optimum HPs by ML algorithms.

Classification algorithm	Optimization method	HP	Search range	Search iterations
RF	Grid search	n_estimators	[100, 110, ..., 990, 1000]	270
		min_samples_split	[7, 13, 20]	
	Random search/ Bayesian optimization	n_estimators	[100–1000]	75
		min_samples_split	[7–20]	
XGBoost	Grid search	max_depth	[2, 4, 6, 8, 10]	3125
		min_child_weight	[1, 2, 3, 4, 5]	
		gamma	[0, 0.1, 0.2, 0.3, 0.4]	
		subsample	[0.6, 0.7, 0.8, 0.9, 1]	
	Random search/ Bayesian optimization	learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]	75
		max_depth	[2–10]	
		min_child_weight	[1–5]	
		gamma	[0–0.4]	
LightGBM	Grid search	subsample	[0.6–1]	625
		learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3]	
		max_depth	[2, 4, 6, 8, 10]	
		min_child_weight	[1, 2, 3, 4, 5]	
	Random search/ Bayesian optimization	subsample	[0.6–1]	75
		learning_rate	[0.01–0.3]	
		max_depth	[2–10]	
		min_child_weight	[1–5]	

utilizes only one subset as validation data, performs model training with $K - 1$ subsets, and is iterated K times.⁽³⁵⁾ The average accuracy of the training and validation datasets divided into K subsets was calculated as the HP accuracy. The HP with the highest average accuracy was selected as the optimum HP.⁽³⁶⁾

3.6 Classification performance and precision evaluation

Test datasets were utilized to evaluate the classification algorithm. To evaluate consistency, a land cover map of the study area was constructed by implementing the optimal model for each algorithm and was compared with the land cover map of the Ministry of Environment. As evaluation measures, the overall accuracy (OA) and kappa coefficient were utilized. The user accuracy (UA, recall) and producer accuracy (PA, precision) were adopted as secondary measures [Eqs. (1)–(6)].^(37–39)

$$\text{Overall Accuracy (OA)} = \frac{TP(\text{True Positive}) + TN(\text{True Negative})}{TP + FP(\text{False Positive}) + TN + FN(\text{False Negative})} \quad (1)$$

$$\text{User Accuracy (UA)} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Producer Accuracy}(PA) = \frac{TP}{TP + FP} \quad (3)$$

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (4)$$

$$p_0 = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$p_e = \frac{(TP + FM) \times (TP + FP) \times (FP + TN) \times (FN + TN)}{(TP + TN + FP + FN)^2} \quad (6)$$

4. Results

4.1 Optimization of HPs in ML algorithms

The OA derived from HP selection decreased in the order XGBoost > LightGBM > RF. XGBoost with the grid search showed the highest performance in terms of OA, and RF with the random search showed the lowest performance. The random search was effective in rapidly finding HPs, while the grid search required a long computational time. The computational time of the grid search method in RF was about twice that of the other optimization methods. The grid search in XGBoost and LightGBM took more than 50 times longer than the random search (Table 3).

4.2 Evaluation of classification performance in ML algorithms

Tables 4–6 show the relationship between the label data and the corresponding classified data for each ML algorithm. The OA and kappa coefficient were 82.87 and 65.91% for the RF model with the grid search, 83.24 and 66.91% for the XGBoost model with the grid search, and 83.08 and 66.73% for the LightGBM model with Bayesian optimization, respectively. The classification performance decreased in the order XGBoost > LightGBM > RF. The accuracy evaluated using the receiver operating characteristic (ROC) curve also decreased in the order XGBoost > LightGBM > RF. In the three ML algorithms, the accuracy of each land cover item evaluated using the ROC curve decreased in the order water > used area > forest > agricultural land > wetland > barren > grass. The three ML algorithms showed no significant difference in accuracy (Fig. 3).

The calculated results of the feature importance exhibited differences between the three models. The feature importance of the RF model decreased in the order Band 4 > Band 3 > Band 5 for satellite images taken during the growth period. For the XGBoost model, the feature

Table 3
Results of HP selection according to the optimization method by algorithm.

Algorithm	Optimization method	HP	Optimal value	OA	Time (min)
RF	Grid search	n_estimators	680	0.8270	58
		min_samples_split	7		
	Random search	n_estimators	678	0.8268	17
		min_samples_split	8		
	Bayesian optimization	n_estimators	898	0.8269	20
		min_samples_split	8		
XGBoost	Grid search	max_depth	10	0.8327	613
		min_child_weight	2		
		gamma	0		
		subsample	0.8		
		learning_rate	0.1		
	Random search	max_depth	8	0.8318	15
		min_child_weight	2		
		gamma	0.38		
		subsample	0.85		
	Bayesian optimization	max_depth	9	0.8313	36
		min_child_weight	2		
		gamma	0.19		
		subsample	0.93		
		learning_rate	0.16		
	Grid search	max_depth	8	0.8305	50
		min_child_weight	5		
		subsample	0.6		
		learning_rate	0.1		
LightGBM	Random search	max_depth	8	0.8303	1
		min_child_weight	4		
		subsample	0.71		
	Bayesian optimization	max_depth	2	0.8279	4
		min_child_weight	3		
		subsample	0.89		
		learning_rate	0.06		

importance decreased in the order Band 3 > Band 4 > Band 2 for satellite images taken during the growth period. For the LightGBM model, the calculated order of the feature importance during the growth period was DEM > Band 8 > Band 2. In addition, the values of feature importance, which was calculated differently for each model, were summed, and the resulting value was assumed to be the relative feature importance of each variable.⁽⁴⁰⁾ The calculated relative feature importance during the growth period was in the order Band 3 > Band 4 > Band 2. In addition, the RGB band was selected as the most important predictor. This result was consistent with the findings of a previous study that reported a high level of separation in the images taken during the growth period in the classification process. This was due to the higher likelihood of separated classification in the images taken during the growth period than those taken during the non-growth period (Fig. 4).

Table 4
RF–label data confusion matrix. (Unit: pixel)

		Label data								
	Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)
RF	Used area	1128	123	71	183	5	189	6	1705	66
	Agricultural land	149	2138	94	467	174	177	20	3219	66
	Forest	25	240	16844	974	35	115	16	18249	92
	Grass	123	435	375	1010	59	202	9	2213	46
	Wetland	—	5	—	15	79	3	3	105	75
	Barren	26	32	33	30	2	122	2	247	49
	Water	4	4	2	—	11	15	220	256	86
	Total	1455	2977	17419	2679	365	823	276	25994	
	PA (%)	78	72	97	38	22	15	80		
	OA (%)						82.87			
	Kappa (%)						65.91			

Table 5
XGBoost–label data confusion matrix. (Unit: pixel)

		Label data								
	Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)
XGBoost	Used area	1107	111	57	174	7	176	4	1636	68
	Agricultural land	135	2129	85	429	129	157	20	3084	69
	Forest	36	215	16799	891	28	102	11	18082	93
	Grass	133	447	431	1112	63	218	10	2414	46
	Wetland	4	29	5	33	122	7	10	210	58
	Barren	39	42	40	39	3	150	3	316	47
	Water	1	4	2	1	13	13	218	252	87
	Total	1455	2977	17419	2679	365	823	276	25994	
	PA (%)	76	72	96	42	33	18	79		
	OA (%)						83.24			
	Kappa (%)						66.91			

Table 6
LightGBM–label data confusion matrix. (Unit: pixel)

		Label data								
	Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)
LightGBM	Used area	1111	116	66	174	7	172	3	1649	67
	Agricultural land	129	2127	89	431	132	162	10	3080	69
	Forest	28	206	16765	885	24	98	8	18014	93
	Grass	128	427	443	1079	51	208	16	2352	46
	Wetland	4	40	2	42	130	11	13	242	54
	Barren	52	57	47	64	3	161	2	386	42
	Water	3	4	7	4	18	11	224	271	83
	Total	1455	2977	17419	2679	365	823	276	25994	
	PA (%)	76	71	96	40	36	20	81		
	OA (%)						83.08			
	Kappa (%)						66.73			

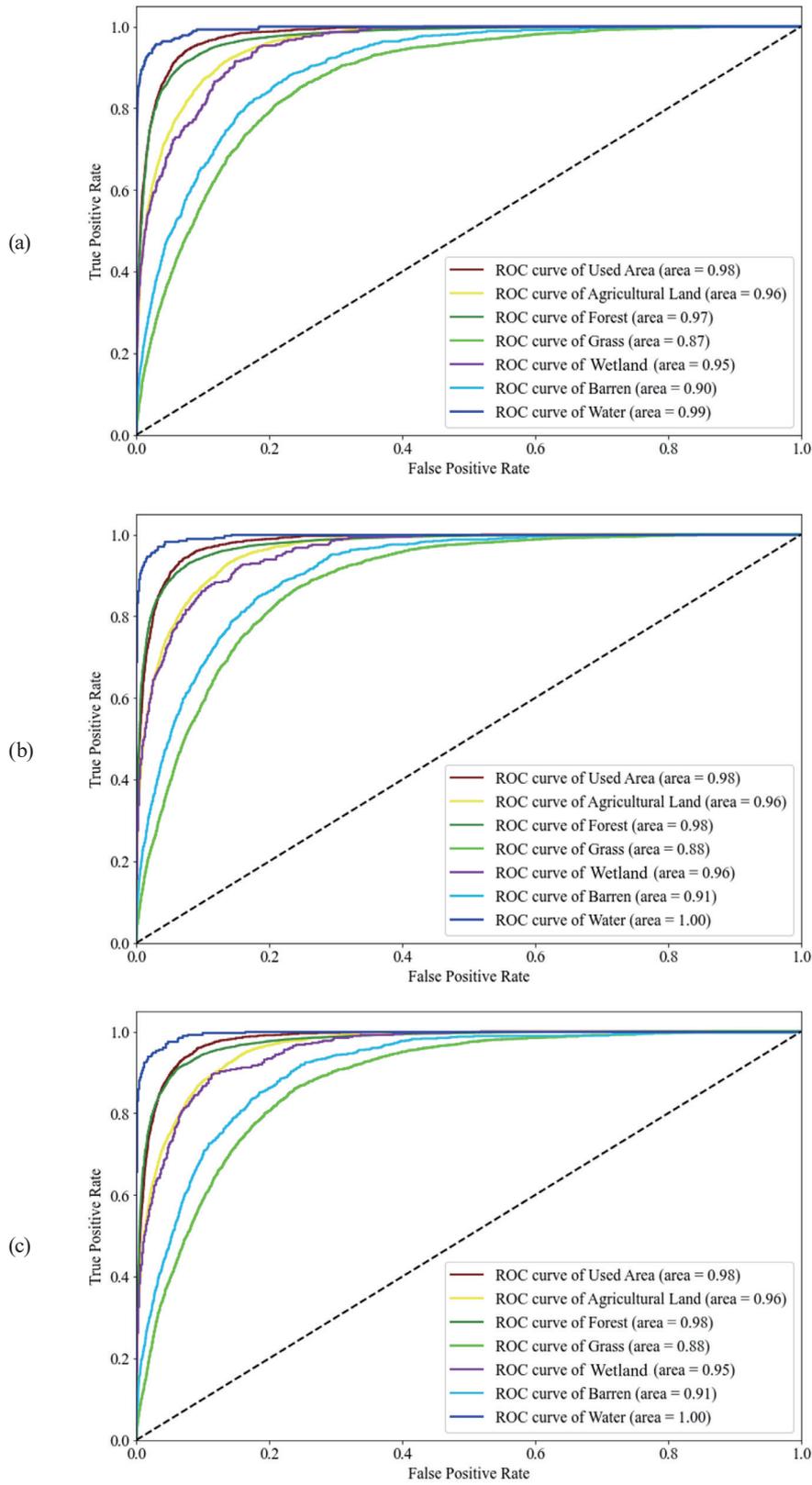


Fig. 3. (Color online) ROC curves: (a) RF, (b) XGBoost, and (c) LightGBM.

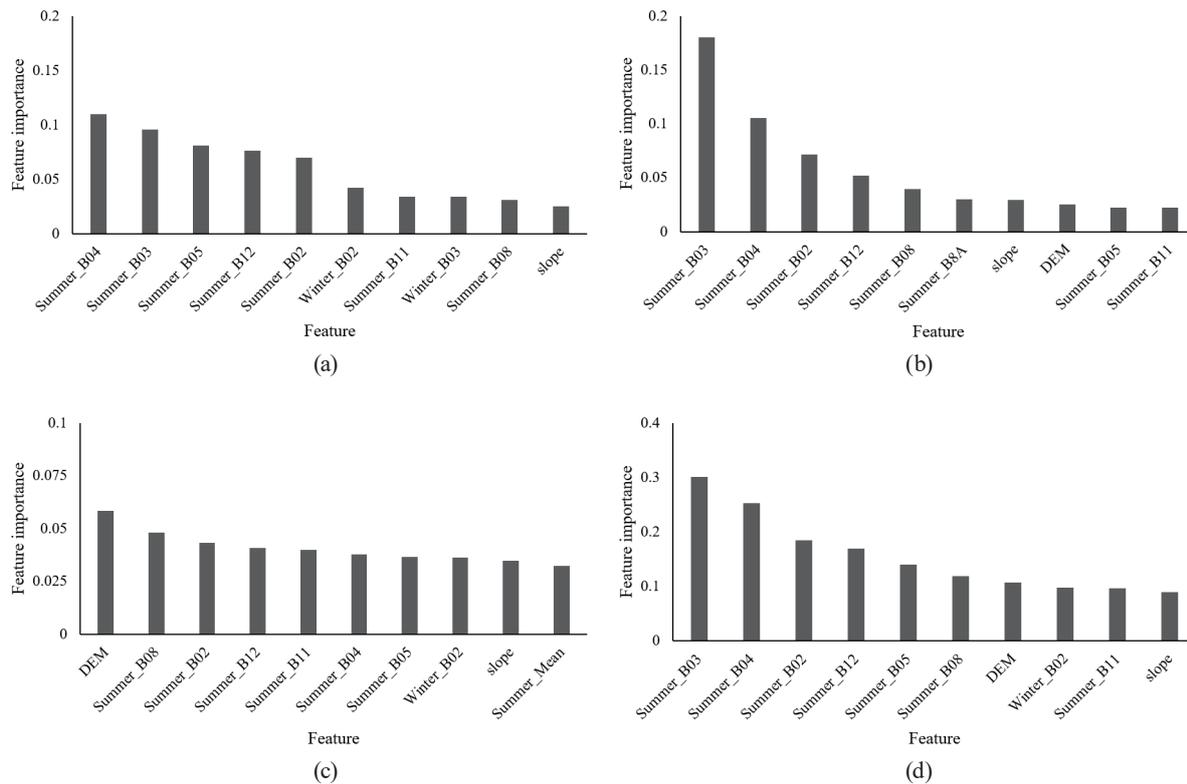


Fig. 4. Ten most important features: (a) RF, (b) XGBoost, (c) LightGBM, and (d) sum of feature importance values.

4.3 Distribution of land cover classification by ML algorithms

Figure 5 shows the land-use land-cover maps obtained by adopting the three land cover classification models. The Ministry of Environment's land cover map classified all narrow streams and roads with a resolution of 1 m. However, the results of ML algorithms that classified maps were generated using satellite images with a spatial resolution of 10 m. Therefore, small streams were classified as grass or agricultural land, and narrow farm roads were classified as agricultural land.

4.4 Precision evaluation with land cover map

We compared the consistency between the land cover maps of the study area constructed using the land cover classification models and the land cover map of the Ministry of Environment. All three models exhibited an OA of approximately 83% and a kappa coefficient of about 66%. The average OA of the RF model was 82.90% and its kappa coefficient was 66.15%, the average OA of the XGBoost model was 83.39% and its kappa coefficient was 67.39%, and the average OA of the LightGBM model was 83.10% and its kappa coefficient was 66.92%. The average OA and kappa coefficient were the best in XGBoost among the three models (Tables 7–9).

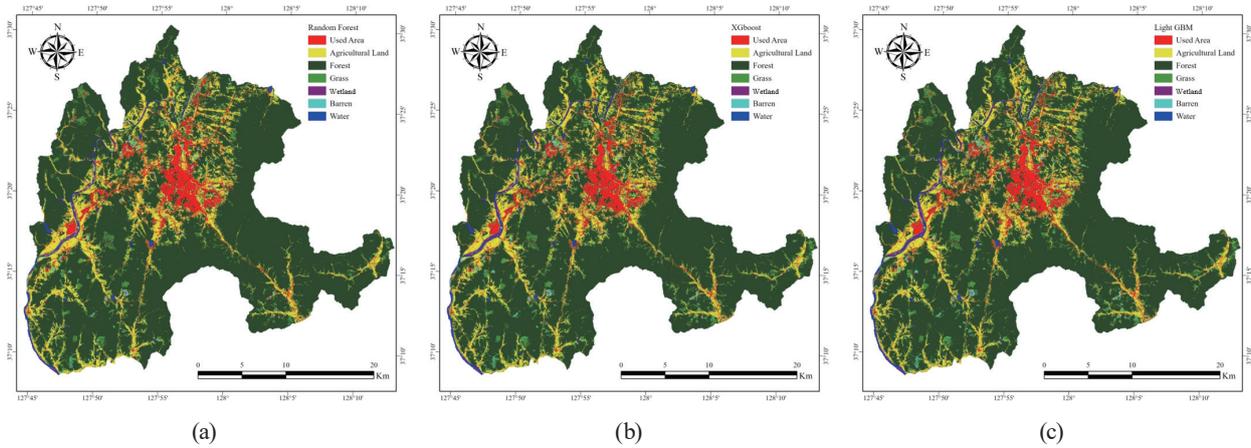


Fig. 5. (Color) Classification maps: (a) RF, (b) XGBoost, and (c) LightGBM.

Table 7
RF–land cover map confusion matrix. (Unit: pixel)

		Label data								
Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)	
Used area	3722	428	223	568	24	627	22	5616	66	
Agricultural land	503	7187	288	1644	577	536	74	10809	66	
Forest	78	709	55977	3316	109	377	37	60604	92	
Grass	407	1366	1291	3484	210	663	46	7467	47	
Wetland	3	26	1	38	292	17	17	392	74	
Barren	120	121	114	103	7	470	2	938	50	
Water	11	18	6	12	46	28	697	818	85	
Total	4845	9855	57900	9166	1265	2718	895	86644		
PA (%)	77	73	97	38	23	17	78			
OA (%)					82.90					
Kappa (%)					66.15					

Table 8
XGBoost–land cover map confusion matrix. (Unit: pixel)

		Label data								
Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)	
Used area	3692	401	191	515	22	564	20	5405	68	
Agricultural land	445	7158	295	1549	409	498	57	10411	69	
Forest	89	605	55874	3021	86	334	26	60033	93	
Grass	445	1442	1406	3822	231	706	50	8101	47	
Wetland	8	78	9	92	455	35	36	714	64	
Barren	158	155	120	157	12	554	6	1162	48	
Water	8	17	7	10	50	26	700	817	86	
Total	4845	9855	57900	9166	1265	2718	895	86644		
PA (%)	76	73	96	42	36	20	78			
OA (%)					83.39					
Kappa (%)					67.39					

Table 9
LightGBM–land cover map confusion matrix. (Unit: pixel)

		Label data								
Type	Used area	Agricultural land	Forest	Grass	Wetland	Barren	Water	Total	UA (%)	
LightGBM	Used area	3718	439	203	556	25	584	21	5546	67
	Agricultural land	444	7091	309	1585	403	497	55	10384	68
	Forest	84	591	55757	2992	78	332	20	59854	93
	Grass	396	1403	1451	3677	201	659	37	7825	47
	Wetland	14	121	18	134	483	50	43	863	56
	Barren	180	185	153	208	16	568	11	1321	43
	Water	9	24	10	14	59	28	707	851	83
	Total	4845	9855	57900	9166	1265	2718	895	86644	
	PA (%)	77	72	96	40	38	21	79		
	OA (%)					83.10				
	Kappa (%)					66.92				

5. Discussion

5.1 HP optimization

The search time of the optimal HPs differed according to the combination of the ML algorithm and the optimization method. The combination of LightGBM and the random search took the shortest time to optimize the parameters, and the combination of XGBoost and the grid search took the longest time. Regardless of the ML model, the time required for the optimization method increased in the order random search < Bayesian optimization < grid search. In addition, the number of searches of the HPs affected the search time in the optimization of the HPs over the search range. In addition, although the number of searches was set to be the same for the random search and Bayesian optimization, the required times for the optimization analysis were highly different owing to the HPs. The combination of XGBoost and Bayesian optimization took the longest time, and the combination of LightGBM and the random search took the shortest time. In particular, the larger the value of the HP $n_estimators$, the longer the time required to build the model. In this study, the random search was superior to Bayesian optimization in terms of analysis time. The random search and Bayesian optimization are characterized by random optimization. Bayesian optimization is a process of determining new parameters by setting the objective function and fitting the Gaussian model, resulting in the time required being longer than that of the random search.

5.2 Model evaluation

In terms of the OA and kappa coefficient, XGBoost was the best model. The kappa coefficient of the matrices for the three ML algorithms was more than 60%, and both the classification results and the label data were consistent. In addition, UA and PA were highest for forest and

lowest for barren. This is because the spectral characteristics of barren are similar to those of fallow agricultural land, construction sites in the used area, and grassland.

In addition, in the classification process using 1-m-resolution images, misclassification occurred because the landscaping areas where herbaceous plants were planted were mixed with label data within one pixel.

5.3 Limitations and future research

In this study, we constructed a dataset using existing spatial data. The advantage of this method is that labeling data can be quickly obtained over a large area. However, even though images produced in the same year as the spatial data were used, the label data were inconsistent for areas that changed rapidly, such as harvested areas in the forest area and construction sites in the used area. Label data were also inconsistent in areas that require a high level of security, such as airports and military bases. Therefore, the supplementation of the data and revisions will be needed to use the spatial data as label data. Furthermore, the spatial resolution of the Sentinel-2 satellite is approximately 10 m. One pixel can thus contain different classes. In the case of pixels in which several categories are mixed, particularly those at class boundaries, the accuracy of pixels can be enhanced by improving the spatial resolution.

6. Conclusions

We applied ensemble ML algorithms based on a decision tree to land cover classification. The dataset used for the ML algorithms was constructed by including topographic information and GLCM values, which were derived through PCA and texture analysis, as well as the spectral information of satellite images. In particular, we presented the process of optimizing HPs, including the range and search iteration of HPs, by applying three methods (grid search, random search, and Bayesian optimization) while tuning the HPs for the application of the ML algorithms. The results of this study are highly applicable to tree-based ML research and can be applied to object-based classification as well as pixel-based classification in the classification of land cover types. It is expected to be possible to classify land cover with high reliability if a deep learning algorithm in the field of CNN-based semantic segmentation is applied. We also constructed a large training dataset using previously established data. If a dataset is constructed by correcting and supplementing data in high-security areas and areas with a time series inconsistency using prepared data in the future, although the cost will increase, it is expected that reliable datasets and models can be built.

Acknowledgments

This study was carried out with the support of the R&D Program for Forest Science Technology (Project No. 2021359B10-2123-BD01) provided by the Korea Forest Service (Korea Forestry Promotion Institute).

References

- 1 S. C. Yu, J. W. An, and J. A. Ok: J. Korea Spatial Inf. Soc. **23** (2015) 3. <https://doi.org/10.12672/ksis.2015.23.3.067>
- 2 J. Perman.: Good Practice Guideline for Land Use, Land-Use Change and Forestry (IPCC, 2003) p. 2.7.
- 3 J. Y. Kim.: J. Sci. Technol. Stu. **17** (2017) 1.
- 4 H. Y. Kim, J. S. Chung, and E. J. Choi: J. Digital Converg. **14** (2016) 2. <https://doi.org/10.14400/JDC.2016.14.2.245>
- 5 S. Raschka, J. Patterson, and C. Nolet: Information **11** (2020) 4. <https://doi.org/10.3390/info11040193>
- 6 P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson: Pattern Recognit. Lett. **27** (2006) 4. <https://doi.org/10.1016/j.patrec.2005.08.011>
- 7 M. M. Hayes, S. N. Miller, and M. A. Murphy: Remote Sens. Lett. **5** (2014) 2. <https://doi.org/10.1080/2150704X.2014.882526>
- 8 S. Georganos, T. Grippa, S. Vanhuyse, M. Lennert, M. Shimoni, and E. Wolff: IEEE Geosci. Remote Sens. Lett. **15** (2018) 4. <https://doi.org/10.1109/LGRS.2018.2803259>
- 9 R. U. Bhagwat and B. U. Shankar: I2CT (2019) <https://doi.org/10.1109/I2CT45611.2019.9033768>
- 10 C. D. Man, T. T. Nguyen, H. Q. Bui, K. Lasko, and T. N. T. Nguyen: Int. J. Remote Sens. **39** (2018) 4. <https://doi.org/10.1080/01431161.2017.1399477>
- 11 P. Carmona, F. Climent, and A. Momparler: Int. Econ. Rev. **61** (2019). <https://doi.org/10.1016/j.iref.2018.03.008>
- 12 T. Zhang, J. Su, Z. Xu, Y. Luo, and J. Li: Appl. Sci. **11** (2021) 2. <https://doi.org/10.3390/app11020543>
- 13 M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, and P. Martimort: Remote Sens. Environ. **120** (2012) <https://doi.org/10.1016/j.rse.2011.11.026>
- 14 A. Agapiou, D. D. Alexakis, A. Sarris, and D. G. Hadjimitsis: Remote Sens. **6** (2014) 3. <https://doi.org/10.3390/rs6032176>
- 15 Ministry of Environment: <https://egis.me.go.kr/intro/land.do> (accessed August 2021).
- 16 T. Gao, J. Zhu, X. Zheng, G. Shang, L. Huang, and S. Wu: Remote Sens. **7** (2015) 2. <https://doi.org/10.3390/rs70201702>
- 17 H. Zheng, P. Du, J. Chen, J. Xia, E. Li, Z. Xu, X. Li, and N. Yokoya: Remote Sens. **9** (2017) 12. <https://doi.org/10.3390/rs9121274>
- 18 P. Kupidura: Remote Sens. **11** (2019) 10. <https://doi.org/10.3390/rs11101233>
- 19 A. K. Saha, M. K. Arora, E. Csaplovics, and R. P. Gupta: Geocarto Int. **20** (2005) 2. <https://doi.org/10.1080/10106040508542343>
- 20 A. M. Abdi: GISci. Remote Sens. **57** (2020) 1. <https://doi.org/10.1080/15481603.2019.1650447>
- 21 S. M. Joy, R. M. Reich, and R. T. Teynolds: Int. J. Remote Sens. **24** (2003) 9. <https://doi.org/10.1080/01431160210154948>
- 22 M. Pal: Int J Remote Sens. **26** (2005) 1. <https://doi.org/10.1080/01431160412331269698>
- 23 Z. Xu, J. Chen, J. Xia, P. Du, H. Zheng, and L. Gan: IEEE Geosci. Remote Sens. Lett. **15** (2018) 5. <https://doi.org/10.1109/LGRS.2018.2806223>
- 24 Q. Bui, T. Chou, T. Hoang, Y. Fang, C. Mu, P. Huang, V. Pham, Q. Nguyen, D. T. N. Anh, and V. Pham: Remote Sens. **13** (2021) 14. <https://doi.org/10.3390/rs13142709>
- 25 M. Belgou, L. Dragut: ISPRS J. Photogramm. Remote Sens. **114** (2016). <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- 26 L. Breiman: Mach Learning **45** (2001) 1
- 27 G. Biau, E. Scornet: TEST **25** (2016) 2. <https://doi.org/10.1007/s11749-016-0481-7>
- 28 T. Chen, C. Guestrin: KDD '16: Proceedings of the 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (2016, KDD). <https://doi.org/10.1145/2939672.2939785>
- 29 H. Jiang, D. Li, W. Jing, J. Xu, J. Huang, J. Yang, and S. Chen: Remote Sens. **11** (2019) 7. <https://doi.org/10.3390/rs11070861>
- 30 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu: Adv. Neural Inf. Process. Syst. **30** (2017).
- 31 A. Y. M. Abdullah, A. Masrur, M. S. G. Adnan, M. Baky, A. Al, Q. K. Hassan, and A. Dewan: Remote Sens. **11** (2019) 7. <https://doi.org/10.3390/rs11070790>
- 32 Y. Li, M. Li, C. Li, and Z. Liu: Sci. Rep. **10** (2020) 1. <https://doi.org/10.1038/s41598-020-67024-3>
- 33 O. D. Kazakov, A. V. Averchenkov, and N. Y. Kulagina: J. Phys. Conf. Ser. **1210** (2019). <https://doi.org/10.1088/1742-6596/1210/1/012061>
- 34 S. Yadav, S. Shukla: 2016 IEEE 6th Int. Conf. Advanced Computing (IEEE, 2016). <https://doi.org/10.1109/IACC.2016.25>

- 35 D. Gómez, P. Salvador, J. Sanz, and J. L. Casanova: Remote Sens. **11** (2019) 15. <https://doi.org/10.3390/rs11151745>
- 36 S. Yadav, S. Shukla: IEEE Trans. Pattern Anal. Mach. Intell. **32** (2010) 3. <https://doi.org/10.1109/TPAMI.2009.187>
- 37 B. Vasilakos, D. Kavroudakis, and A. Georganta: Remote Sens. **12** (2020) 12. <https://doi.org/10.3390/rs12122005>
- 38 C. Liu, P. Frazier, and L. Kumar: Remote Sens. Environ. **107** (2007) 4. <https://doi.org/10.1016/j.rse.2006.10.010>
- 39 S. Stehman: Photogramm Eng. Remote Sens. **62** (1996) 4.
- 40 E. Akbari, A. D. Boloorani, N. N. Samany, S. Hamzeh, S. Soufizadeh, and S. Pignatti: Remote Sens. **12** (2020) 9. <https://doi.org/10.3390/rs12091449>

About the Authors



Jeongmook Park received his B.S., M.S., and Ph.D. degrees from Kangwon National University, Korea, in 2011, 2013, and 2019, respectively. Since 2019, he has been a postdoctoral researcher at Kangwon National University. His research interests include forest spatial analysis using GIS and RS. (jm_park@kangwon.ac.kr)



Yongkyu Lee received his B.S. and M.S. degrees from Kangwon National University, Korea, in 2019 and 2021, respectively, and is currently studying for a Ph.D. degree. His research interests include forest spatial analysis using GIS and RS. (kong445566@kangwon.ac.kr)



Jungsoo Lee received his B.S. degree from Kangwon National University, Republic of Korea, in 1998 and his M.S. and Ph.D. degrees from Tokyo University, Japan, in 2000 and 2003, respectively. Since 2006, he has been a professor at Kangwon National University. His research interests include GIS and remote sensing related to forest management. (jslee72@kangwon.ac.kr)