TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 24, No. 4, August 2026, pp. 1307~1319
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i4.27747

Decision-tree-based machine learning for detecting coffee agroforestry using SPOT-7

I Made Khrisna Yoga Devandra, I Nengah Surati Jaya, Tatang Tiryana
Department of Forest Management, Faculty of Forestry and Environment, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Dec 20, 2025
Revised Apr 7, 2026
Accepted May 25, 2026

Keywords:

Biophysical variables
Coffee agroforestry
Decision tree
Machine learning
Spectral variables
SPOT-7 image

ABSTRACT

This study develops a decision-tree-based machine-learning (ML) approach to identify coffee agroforestry plants using SPOT-7 satellite imagery. The algorithm was developed by examining the combination of image indices derived from SPOT-7 and biophysical variables. Detection using spectral variables is often hampered by spectral similarity between vegetation cover classes. This study found that a ML method that combines spectral and biophysical variables can significantly improve overall accuracy, from 60.4% (using conventional spectral variables alone) to 94% (using integrated spectral-biophysical variables). For detecting and identifying agroforestry coffee classes typically found under tree canopies, the addition of the "land cover" variable published by the Ministry of Environment and Forestry contributes significantly to the classification of agroforestry coffee. Important variables identified in this model are normalized difference vegetation index (NDVI), visible difference vegetation index (VDVI), normalized red-green vegetation index (NRGI), elevation, and land cover.

Corresponding Author:

I Nengah Surati Jaya
Department of Forest Management, Faculty of Forestry and Environment, IPB University
Bogor (16680), Indonesia
Email: ins-jaya@apps.ipb.ac.id

1. INTRODUCTION

Over the last two decades, agroforestry has emerged as a strategic approach to sustainable forest resource management within the framework of multi-business forestry, also known as "multi-usaha kehutanan (MUK)" in Indonesia. This multi-business system has integrated forestry and agricultural commodities under a single management unit, enabling simultaneous ecological and economic benefits. It has been shown that agroforestry practices enhance land productivity, both in terms of forest cover utilization and agricultural yield [1].

Tanggamus Regency in Lampung Province is a good example of agroforestry coffee practices, with an extensive forest area of approximately 1,476.69 km², including protected forests and conservation areas. At this study site, Robusta coffee is the dominant species, playing a crucial role in economic value and ecological adaptability. This species demonstrates a broad ecological tolerance, growing across a wide altitudinal range under shaded conditions [2], [3].

The presence of shade trees is essential in regulating microclimatic conditions, particularly by reducing excessive radiation exposure and improving plant productivity. Consequently, coffee-based agroforestry systems have become an important component of both local livelihoods and global agricultural markets [1], [3]. Coffee grown in agroforestry systems makes a significant contribution to community well-being while also generating positive ecological impacts. However, accurate data on the extent of these agroforestry coffee plantations is extremely difficult to obtain. The problem is that identifying and mapping

the spatial distribution of coffee agroforestry systems, especially those located beneath the forest canopy, is extremely difficult. Terrestrial mapping or interviews are inefficient inventory methods for large-scale applications. Furthermore, the availability of optical remote sensing data is limited to only objects exposed at the top of the tree canopy or those captured by cameras. Misclassification often occurs among canopy cover classes when similar classes are spectrally mixed, due to spectral mixing between coffee plants and other canopy vegetation. Therefore, reliable spatial information on the extent and distribution of coffee agroforestry systems is severely limited, hampering effective planning, monitoring, and policymaking.

Recent advances in remote sensing and machine learning (ML) have provided new opportunities for land cover classification. Among these, decision tree algorithms have gained attention for their simplicity, interpretability, and reproducibility across diverse datasets [4]–[6]. Previous studies have demonstrated that integrating spectral variables derived from satellite imagery with non-spectral variables, such as biophysical and geographical factors, can significantly improve classification accuracy. However, applying these approaches to complex agroforestry systems, particularly for detecting understory crops such as coffee, remains challenging.

Previous research has identified several critical issues, particularly in the use of variables derived solely from spectral data [4], [7]–[9]. First, the effectiveness of spectral image-based decision-tree algorithms for detecting understory coffee agroforests remains low because optical imagery provides only spectral information about the vegetation canopy [10]. Second, understanding the most influential variables in determining coffee agroforestry locations is strongly influenced by site suitability (biophysical factors) and socioeconomic factors (road access, distance from settlement centers) [11]. Third, challenges related to model generalization, including overfitting and transferability, remain insufficiently addressed [12]. These conditions create uncertainty in model development and limit the operational application of ML approaches for agroforestry mapping in tropical regions.

To address these challenges, the objective of this study is to develop a decision tree-based ML model that integrates spectral variables derived from SPOT-7 imagery with non-spectral variables such as biophysical and socio-economic variables in Tanggamus Regency. Specifically, this study seeks to identify key variables and their relative importance, evaluate the model's performance and reliability, and assess its interpretability. This study is expected to improve the accuracy and applicability of spatial mapping for coffee agroforestry systems. The developed model is expected to provide a robust and easily understood framework to support sustainable land management and agroforestry monitoring in tropical environments.

2. METHOD

Tanggamus Regency is located at 104°18'–105°12' E and 5°05'–5°56' S, encompassing 4654.95 km² with elevations reaching 2115 m (see Figure 1). Approximately 40% of the region comprises plateau landscapes, including hilly to mountainous terrain. Tanggamus Regency has a coffee plantation area of 41,512 hectares and a production of 33,482 tons. For on-the-ground observation and data processing and analysis, we used a global positioning system (GPS) device, specific software such as ArcGIS and ERDAS Imagine for spatial analysis and statistical data extraction, and RapidMiner Studio for developing the decision tree algorithm. The algorithm was then implemented in Python.

The main data used are SPOT-7 imagery, which includes one panchromatic (PAN) band with a high spatial resolution of 1.5 m and four multispectral bands spanning 0.450 to 0.745 μm, enabling detailed spatial observation. The multispectral bands are provided at a spatial resolution of 6 m, including three visible bands (blue, green, and red) and a near-infrared band spanning 0.760–0.890 μm. These spectral bands collectively support a wide range of applications, including vegetation analysis, land cover classification, and environmental monitoring [13]. These bands were then converted into several indices, namely the normalized difference vegetation index (NDVI), the normalized red-green vegetation index (NRGI), the visible difference vegetation index (VDVI), and the visible atmosphere resistance (VARI). These indices leverage spectral data to provide insights into vegetation dynamics, offering a non-destructive means of monitoring and managing forest resources, particularly for assessing vegetation classes in dense-canopy agroforestry plantations. Each index has unique strengths: NDVI is used to assess vegetation health due to its sensitivity to chlorophyll content and canopy structure, and NRGI, which is similar to the Red-Edge NDVI, has potential for biomass estimation, particularly when combined with other indices [14]. The VDVI, akin to VARI, is designed to be less sensitive to atmospheric conditions, making it useful for monitoring vegetation health in variable weather conditions [10]. VARI is effective in distinguishing vegetation from soil and other non-vegetative surfaces, which is crucial in dense canopy environments where ground cover can obscure vegetation signals [15].

To improve model performance, the study also evaluated the contribution of biophysical variables such as elevation, slope, visual common land cover, and proximity to rivers, roads, and settlements. Elevation

and slope were derived from National Aeronautics and Space Administration digital elevation model (NASADEM), while visual land cover was interpreted from SPOT-7 imagery. Furthermore, proximity-based variables were incorporated, including distances to rivers, roads, and settlements, obtained from the *Rupa Bumi Indonesia* (RBI) map. Field-based sample plots collected through ground surveys in 2023 were used as reference data, while administrative boundaries (district and subdistrict levels) were also sourced from RBI 2020 data to support spatial analysis.

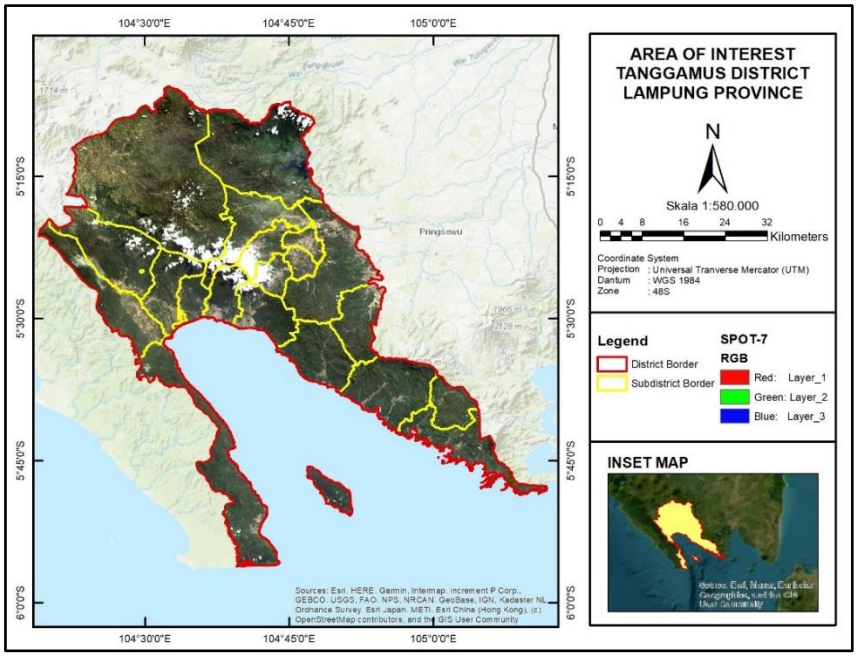


Figure 1. Research location in Tanggamus District

2.1. Image preprocessing

Prior to analysis, the spatial data were geo-corrected and preprocessed, including index development, spatial filtering, and image [16]. Radiometric correction was performed to improve image quality by mitigating surface reflectance errors and atmospheric influences, obtaining more accurate information [17].

2.2. Class scheme and training area development

The class scheme was developed through visual interpretation of imagery based on elements such as shape, color/hue, pattern, texture, shadow size, location, and association [17]. The class scheme was divided into 10 classes: water bodies (WTB), plantations/estate crop (PLT), scrub (SCB), natural forests (FRS), bare land (BRL), settlements (STL), mixed dryland agriculture (AGC), paddy fields (PDF), coffee agroforestry (CAF), and non-cloud land cover (CLD). A training area of 6,666 points was created based on the class scheme and distributed evenly across land cover. The location of the training area was determined based on image appearance and field surveys of the coffee class, both agroforestry and monoculture. The study used a training area with a 2.25 m buffer, or 4.5 m × 4.5 m (20.25 m^2). We used a sample size of 4.5×4.5 (or a 2.25 m point buffer) because the image pixel size is 1.5 m. This size allows for representative data, with one observation point covering 9 pixels (3×3 pixels).

2.3. Decision tree algorithm development

The decision tree algorithm is a non-parametric method capable of handling large and complex datasets without requiring a complex parametric structure. Figure 2 shows the procedure for developing the decision tree algorithm of this study [4], [18]. The data were divided into a training dataset (70%) and a validation dataset (30%). The training dataset was used to build a decision tree model, and the validation dataset was used to determine the optimal tree size to achieve the final model [19]. The importance of variables was determined using the Brute Force method (1), information gain (2), gain ratio (3), and gini index (4) using the formulas as depicted from (1) to (6) [18].

$$Entropy(S) = -\sum_{i=1}^{c} p_i \, log_2(p_i) \quad (1)$$

$$IG(S,A) = Entropy(S) - \sum_{v\hat{\imath} \, Values(A)}^{n} \frac{|S_v|}{|S|} Entropy\,(S_v) \quad (2)$$

$$Gain\ Ratio\ (S,A) = \frac{IG(S,A)}{Split\ Informatin(A)} \quad (3)$$

$$Split\ Information\ (A) = -\sum_{v\hat{\imath} \, Values(A)}^{n} \frac{|S_v|}{|S|} log_2\left(\frac{|S_v|}{|S|}\right) \quad (4)$$

$$Gini\ (S) = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (5)$$

$$Gini\ Indeks\ (S,A) = \sum_{v\hat{\imath} \, Values(A)}^{n} \frac{|S_v|}{|S|} Gini\ (S_v) \quad (6)$$

Furthermore, the model performance evaluation was conducted by measuring overall accuracy and kappa accuracy using (7), (8) and (9) [12], [14], [15], [20], [21]. The relevance of this evaluation approach is supported by recent studies such as Nuthammachot and Stratoulias [20], who successfully applied vegetation indices derived from Sentinel-2 satellite imagery to classify healthy and diseased oil palm trees, thereby demonstrating the applicability of decision tree-based classification in agricultural contexts.

$$Overall\ Accuarcy\ (OA) = \frac{\sum_{i=1}^{k} x_{ii}}{N} 100\% \quad (7)$$

$$Kappa\ accuracy\ (K) = \frac{OA - P_e}{1 - P_e} \quad (8)$$

$$P_e = \frac{\sum_{i=1}^{k} (x_{i+} . x_{+i})}{N^2} \quad (9)$$

Where S is the data set (sample data) beiing tested, p_i is the proportion of the i-th class, c is the number of classes. S_v is the number of data points at value v for each variable A, where v (sub-attribute) is a member of attribute A. A is the attribute or variable being tested, p_i is the proportion of the i-th class in the total data set $S, IG\ (S,A)$ is information gain for variable A from the entire data set (S). A higher IG value for a variable indicates better class separation in the decision tree.

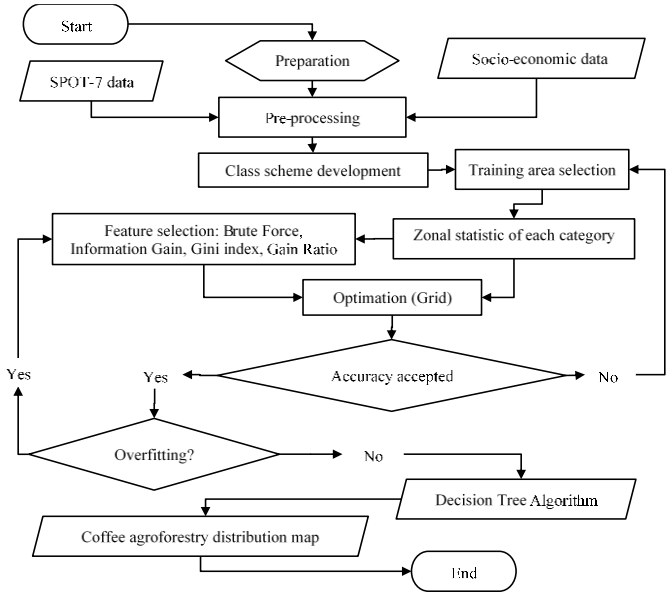


Figure 2. ML-based models and experimental methods applied

3. RESULTS AND DISCUSSION

3.1. Feature selection

The results of a study using a full feature set of 10 variables showed significant overfitting, resulting in a somewhat odd and unnatural spatial map. Accuracy was high, but the visual classification results resembled the shapes of the input variables. Therefore, this study selected variables to reduce overfitting, thereby enabling optimal model performance [9], [21].

The large land cover variable values in both the overfitting and best algorithm results may be influenced by land cover variable data generated from visual interpretation of satellite imagery and by the creation of training area data tailored to the same land cover class. These factors can significantly enhance classification accuracy. In 2022, as shown in Table 1, rainfall in Tanggamus Regency was 154 mm per month, a moderate rainfall category. A weight of 0 indicates that coffee farmers in Tanggamus Regency tend not to rely on river water but instead on rainfall [9]. A variable with a value of 1 indicates that the variable plays a dominant role in the algorithm. Conversely, a value close to 0 indicates that the attribute is not very significant in reducing entropy. A value of zero indicates its lowest importance and thus is subordinate. This study found important variables that can significantly improve classification accuracy. In 2022, it was reported that rainfall in Tanggamus Regency was 154 mm per month, which falls within the moderate rainfall category. A weight of 0 indicates that coffee farmers in Tanggamus Regency tend not to rely on river water but on rainfall.

Table 1. Weight results for each variable (overfitting)

Brute force		Information gain		Gain ratio		Gini index	
Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight
Land cover	1	Land cover	1	Settlement	1	Land cover	1
Elevation	1	NDVI	0.37	NDVI	0.90	Settlement	0.32
Slope	1	NRGI	0.34	Land cover	0.82	NDVI	0.22
River	1	VARI	0.33	NRGI	0.72	NRGI	0.20
Road	1	Settlement	0.29	VARI	0.70	VARI	0.20
Settlement	1	Slope	0.20	Slope	0.33	Road	0.16
NDVI	1	VDVI	0.19	Road	0.32	Slope	0.13
NRGI	1	Elevation	0.17	VDVI	0.31	VDVI	0.13
VDVI	1	Road	0.13	Elevation	0.30	Elevation	0.10
VARI	1	River	0	River	0	River	0

Based on this information, variables such as settlements exhibit inconsistent weight sequences, indicating overfitting, as they yield more branch nodes than spectral variables, leading to illogical classification results. Overfitting is a common problem in machine learning, especially with decision trees, and cannot be avoided entirely [9]. This can be due to limitations in training data, which can be limited in size or encompass many variables, or to limitations in algorithms that are too complex and require too many parameters.

Variables such as road proximity, settlements, rivers, and slopes were removed using the dropout method as in Table 2. This method significantly reduces computational effort and is suitable for large and complex data sets [9]. Land cover received the highest weighting for almost every parameter. Land cover carries a high weighting because it reflects the social and natural conditions of a region and plays a crucial role in understanding the complex relationships between activities and changes occurring on the Earth's surface. Therefore, it has high sensitivity and is the most influential in classifying land cover [11]. The spectral variable with the highest weighting among the best features was NDVI, with a gain ratio of 1, a Gini index of 0.11, and an information gain of 0.25 (see Table 2). This indicates that the NDVI vegetation index is highly reliable in land cover classification. NDVI reflects ecological information, including vegetation growth, vegetation cover level, and biomass [22].

Table 2. Weight result for each variable (best)

Brute force		Information gain		Gain ratio		Gini index	
Variable	Weight	Variable	Weight	Variable	Weight	Variable	Weight
Land cover	1	Land cover	1	NDVI	1	Land cover	1
Elevation	1	NDVI	0.25	Land cover	0.97	NDVI	0.11
NDVI	1	VARI	0.14	VARI	0.59	VARI	0.07
NRGI	1	NRGI	0.13	NRGI	0.57	NRGI	0.06
VDVI	1	Elevation	0.00	Elevation	0.05	VDVI	0.02
VARI	1	VDVI	0	VDVI	0	Elevation	0

3.2. Model optimization

Parameter tuning in decision trees uses grid search, a systematic process that tests various combinations of parameter values to automatically find the optimal one [23]. Model (grid) optimization is an operation that finds the optimal value for each selected parameter, thereby determining the value that yields the maximum results without having to re-enter data. Table 3 shows that the information gain parameter achieved the highest accuracy of 94.0% with the treatment without pruning and without pre-pruning, a maximum depth of 35, a minimum leaf size of 60, a minimum split size of 48, and a pre-pruning alternative of 60.

The results of model optimization using the information gain parameter dominated the parameter space, with the 10 best model parameter combinations out of 10,368. Based on the model optimization, the highest accuracy was 94.0%, demonstrating the decision tree model's ability to classify agroforestry coffee with high accuracy. Information gain proved to be the most effective separation parameter in this case, yielding consistently high accuracy. Information gain is a suitable metric for selecting data splits [5], [16], [23].

Table 3. 10 highest accuracy parameter combination

Parameter	Prunning	Maximal depth	Pre-prunning	Minimal leaf size	Minimal size for split	Pre-prunning alternative	Accuracy (%)
IG	F	35	F	60	48	60	94.0
IG	F	24	F	80	80	0	93.9
IG	T	35	F	1	64	100	93.7
IG	T	24	F	21	17	40	93.7
IG	F	35	F	41	1	40	93.6
GI	F	35	F	21	80	40	93.6
IG	F	24	F	21	17	80	93.6
IG	T	30	F	80	48	100	93.6
IG	F	35	F	41	1	20	93.5
IG	F	19	F	80	48	20	93.5

*Note: GI = gini index, IG = information gain, F = false, T = true

The accuracy-class frequency results for the decision tree parameters indicate that information gain was selected as the best parameter for building a decision tree model, with 10368 combinations achieving 90%-95% accuracy, as shown in Table 4. Gain ratio has more combinations with accuracy below 70% than other parameters, namely 1379. This shows that the gain ratio does not perform well in decision-tree classification. Similar results were reported by [4], who identified information gain as one of the 10 best parameters in the optimization model, achieving an accuracy of 84.6%. The gini index was also selected, achieving the highest accuracy of 84.2%. The roles of the gini index and information gain were frequently recognized in the use of the decision tree algorithm [18].

Table 4. Class accuracy frequency for each decision tree parameter

Class accuracy	Parameter			Total
	Information gain	Gain ratio	Gini index	
90–95%	1727	1697	1709	5133
85–90%	1	326	19	346
80–85%	0	26	0	26
75–80%	1722	19	1728	3469
70–75%	6	9	0	15
<70%	0	1379	0	1379
Total	3456	3456	3456	10368

3.3. Accuracy assessment

The accuracy of training-area data in the decision-tree classification algorithm is assessed using producer, user, total, and kappa accuracies. Producer accuracy and user accuracy are components of the overall accuracy estimate that describe the total accuracy of the classification results; kappa accuracy, in contrast, is determined not only by correctly classified objects but also by classification errors [19]. Three different combinations used in the test are: (i) spectral, bio-geophysical, (ii) spectral, and (iii) bio-geophysical. The accuracy comparison for each combination is shown in Figure 3. Data in the form of a confusion matrix cannot be assessed solely based on numbers and existing accuracy. However, overfitting is evident in the model's visualizations [8], [9], [24].

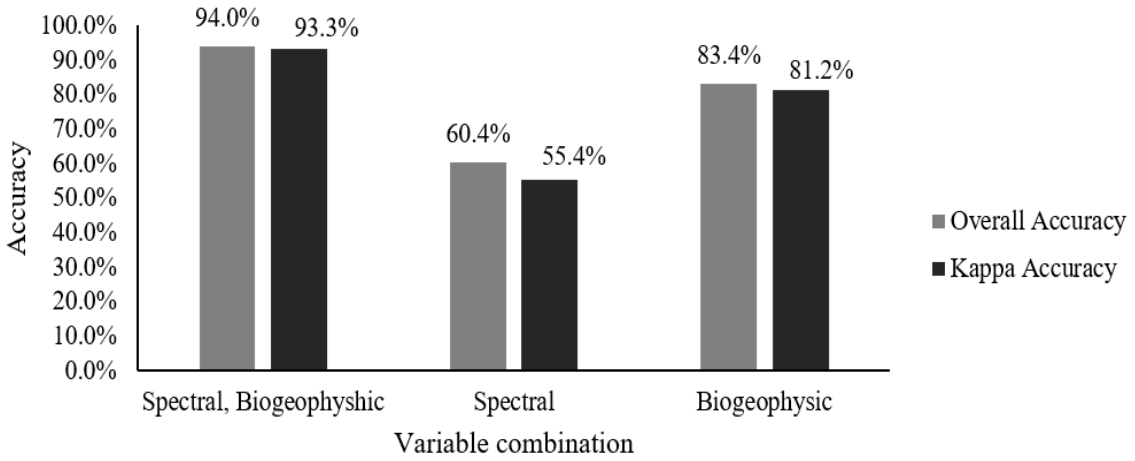


Figure 3. The accuracy difference between each variable combination

The best accuracy was achieved using a combination of spectral and biophysical variables, with an overall accuracy of 94.0% and a kappa of 93.3%. The lowest accuracy was achieved with only spectral variables, with an overall accuracy of 60.4% and a kappa of 55.4%. Spectral variables alone are insufficient to classify agroforestry coffee under a tree canopy. Shading can influence the vegetation index value in coffee fields, so the value obtained does not represent the entire coffee plant vegetation index but rather a mixture of coffee plant and shade tree vegetation indices [14], [15], [20], [25].

The overall accuracy value of each land cover class is a good value (see Table 5), which is in line with [24], which states that accuracy values in the range of 78.8%–100% are included in the medium-high category, meaning the accuracy of the classification results for all land cover classes produced can be fully trusted. Coffee agroforestry has the lowest user accuracy, at 83.6%, indicating an 83.6% chance that the training pixels classified represent agroforestry coffee. However, dryland agriculture has the lowest producer accuracy of 89.2%. This indicates that only 89.2% of the dryland agricultural class training data matches field conditions. The study also showed that the F1-score for each class exceeded 85%.

Table 5. The accuracy results from each variable

Land cover class	User accuracy (recall)	Producer accuracy (precision)	F1 score	Overall accuracy	Kappa accuracy
Cloud	96.8%	97.7%	97.3%	94.0%	93.3%
Coffee agroforestry	83.6%	89.5%	86.5%		
Forest	95.1%	96.1%	95.6%		
Plantation (estate crop)	99.0%	98.1%	98.6%		
Agriculture (dry land agriculture)	92.3%	89.2%	90.7%		
Settlement	97.6%	95.3%	96.4%		
Scrub	92.7%	89.8%	91.2%		
Paddy field	94.4%	99.2%	96.7%		
Water body	99.2%	96.7%	97.9%		
Bare land	97.8%	97.8%	97.8%		
Average	94.8%	94.9%	94.9%		

Similar results were also found in [26], who obtained producer accuracy for one class that did not reach 85%, but achieved an overall accuracy of 94.7% and a kappa coefficient of 0.9, indicating that the level of classification suitability is very high compared to random classification. The most common classification errors found in the dryland agricultural class are caused by the location of dryland agriculture, especially those located or associated with agroforestry coffee and shrub classes, so that they may have vegetation index values that are not much different.

The decision tree algorithm tends to achieve higher accuracy. This is in line with the results of research by [6], which shows the superiority of random forest (an ensemble of decision-tree algorithms) over SVM in classifying urban and forest areas [4], [5], [27]. In identifying coniferous forest classes, the SVM algorithm achieves 95.8% accuracy, but accuracy is lower for other forest types, reaching only 88.3%. Although ML algorithms often promise better accuracy, challenges remain, especially in diverse forest types, particularly in mixed vegetation environments, where spectral variation is very low. This aspect encourages

researchers to continue studying to select the optimal algorithm and select the most consistent attributes to achieve high accuracy.

The resulting decision tree model can classify agroforestry coffee from other land cover types, with a root node representing visual land cover in Figure 4, ELEV = elevation, PLT = plantation (estate crop), and PLVIS = visual land cover. Theoretically, it indicates that agroforestry coffee plants are found in certain land-use types, such as dryland agriculture, plantation forests, and natural forests. Agroforestry plants are generally planted in dry areas because coffee plants are intolerant of waterlogging [1], [3]. Furthermore, coffee plants are grown as a short-term savings crop for farmers, with productive harvests occurring approximately every two months. Coffee plants require good drainage, and this is often found in dryland agriculture. Furthermore, coffee plants are among the most studied plants. Coffee plants typically thrive on mountain slopes and forested areas. By nature, coffee is a dryland plant.

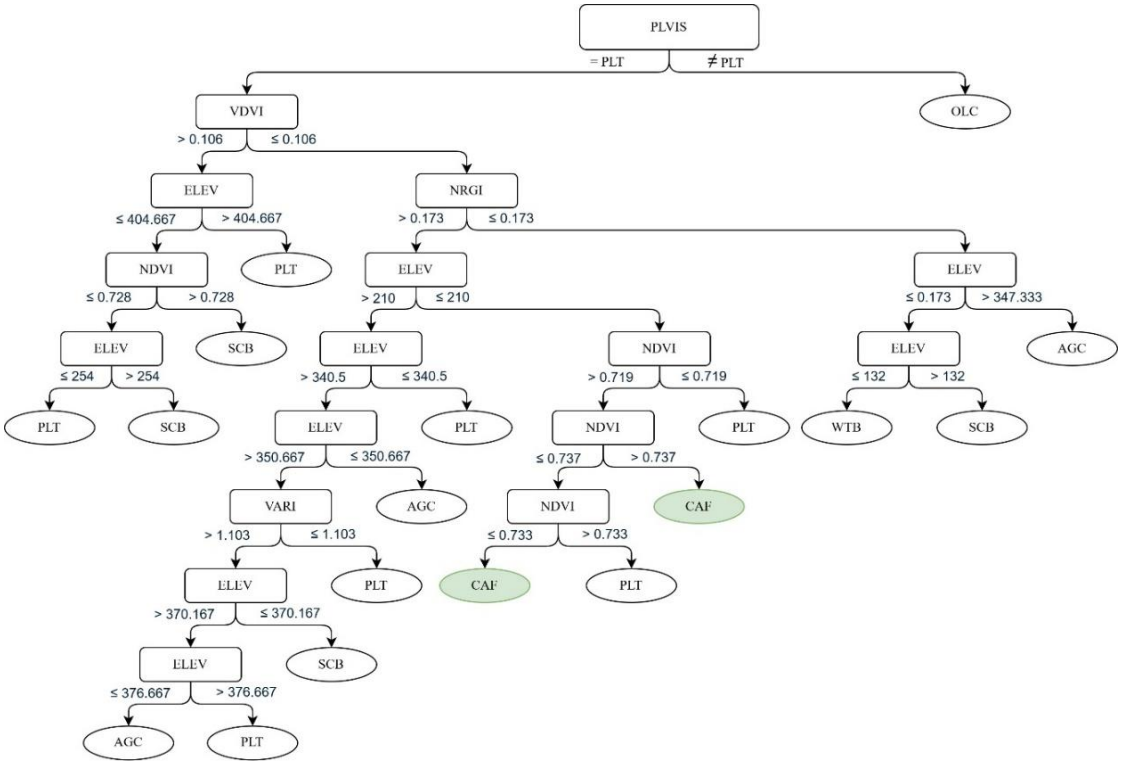


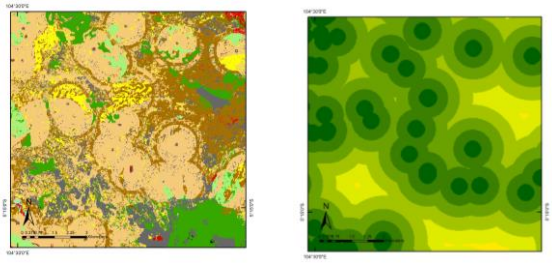
Figure 4. Decision tree model used for land cover classification

Figure 4 shows that the decision tree model relies on a combination of spectral and geophysical variables, which are crucial for accurately distinguishing land cover classes. The decision tree diagram provides context for understanding how various environmental predictors interact and are sequentially selected by the ML algorithm to distinguish different land cover types. In this model, agroforestry coffee was selected based on VDVI ≤ 0.106, NRGI ≤ 0.173, Elevation ≤ 210, and NDVI between 0.719 and 0.737 [3] stated that robusta coffee can be planted in lowland areas with an elevation of less than 800 meters above sea level. NDVI value for coffee agroforestry is also selected around 0.7, indicating the coffee plantation is covered with tree shade, causing the high NDVI value similar to [4] who states that the NDVI value for Tropical Rainforest is above 0.5 and close to 1, and that the greenery/vegetation value is around 0.1–0.4. This is in line with previous research regarding the significant role of NDVI in detecting vegetation variation [28].

3.4. Classification model

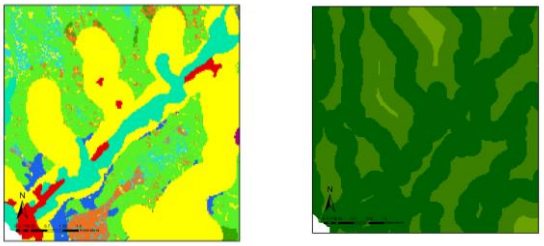
The impact of overfitting on the classification results is clearly shown in Figures 5(a) and (b) to Figures 8(a) and (b); the classification results exhibit a strange pattern that follows the input variable's shape (roads, rivers, settlements, slopes). This indicates that the model is overly sensitive to slight variations in these variables, limiting its ability to distinguish among land cover classes. Overfitting occurs when a model achieves high accuracy on the training data but fails to generalize well across multiple datasets [8], [19]. If

the data is overfit, it will have an unnatural shape. Data comparisons are performed to visualize anomalies. In some cases, overfitting occurs when the algorithm overlearns a single attribute. This is caused by a limited training dataset, an imbalance in the training dataset, and model complexity [10], [29].



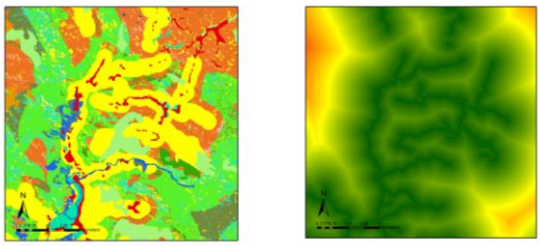
(a) (b)

Figure 5. Overfitting in estate crop classification: (a) settlement-like estate crop pattern and (b) settlement proximity pattern



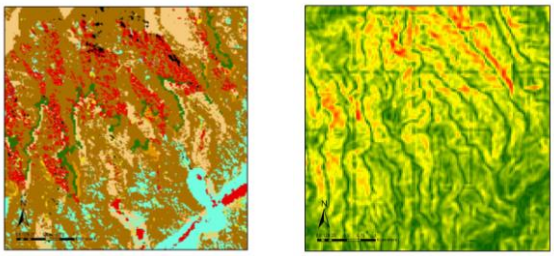
(a) (b)

Figure 6. Overfitting in agricultural classification: (a) river-proximity-like agricultural pattern and (b) river proximity pattern



(a) (b)

Figure 7. Overfitting in scrub classification: (a) road-proximity-like scrub pattern and (b) road proximity pattern



(a) (b)

Figure 8. Overfitting in settlement classification: (a) slope-like settlement pattern and (b) slope pattern

Although the classification results demonstrate relatively high accuracy, the proposed method still exhibits several limitations. This high accuracy fails to detect weaknesses in ML approaches, such as overfitting. Therefore, the main limitation of this method is the lack of a standard method for measuring model reliability. Overall accuracy, the kappa coefficient, recall, and precision are reliable statistical measures for assessing accuracy. However, they still have limitations in measuring spatial errors, including proximity and overfitting.

In this study, because overfitting patterns cannot be identified solely with numerical metrics, a visual approach was used to diagnose them [8], [9]. However, this approach is still entirely qualitative and lacks standardization as a method for diagnosing errors in spatial machine learning. Therefore, the approach relies heavily on the analyst's knowledge and skills (expert judgment), which are highly subjective. Therefore, further research is needed to develop a standardized spatial diagnostic method that integrates spatial patterns and accuracy metrics. In this study, to reduce overfitting, we applied a regularization strategy [30] that includes (1) dropping the attributes that cause overfitting by eliminating the dependence of the model on that specific feature, (2) applying pre-pruning, post-pruning, setting the minimum class members before splitting, and adjusting the tree depth. The first strategy is implemented because the number of point samples is limited [8]. It is very difficult to increase the number of large ground-survey samples, so our main focus is to eliminate variables that lead to severe overfitting.

As shown in Figures 5 to 8, the settlements, roads, rivers, and slopes variables strongly contribute to the overfitting. Therefore, these variables are not used in the model. The variables retained are the image indices (NDVI, NRGI, ARVI, and VDVI), the elevation variables, and the visual vegetation cover variables. The second strategy is implemented to eliminate noise from samples that do not represent field conditions. The pre-pruning (early stopping) limits tree growth during the search for maximum information gain by enforcing constraints such as maximum tree depth and, importantly, the minimum number of samples required for node splitting and for terminal leaf nodes [4]. By increasing the minimum number of samples for splits and leaves, the model avoids creating overly specific branches that are sensitive to small fluctuations in

the data, thereby reducing variance and improving generalization. Post-pruning (pruning), in contrast, allows the decision tree to fully grow before systematically removing branches that contribute little to predictive performance. The focus on performance (overfitting) and interpretability (decision tree) is vital for fostering trust and facilitating informed decision-making in critical sectors where AI is increasingly deployed [31]. Ultimately, addressing overfitting while enhancing model interpretability can lead to more reliable and trustworthy AI systems, fostering greater acceptance and effective use in critical decision-making contexts.

The resulting classification map for the 10 land cover classes is more logical and natural, and does not reveal any trend patterns for any particular variable as in Figure 9. The agroforestry coffee land cover resulting from the algorithmic model classification is shown in dark brown. In general, Tanggamus Regency is dominated by dryland agricultural land cover. In accordance with the objectives of this study, the land cover classification map aims to facilitate the visualization of land cover results. This visualization helps depict an area for predicting land cover and detecting agroforestry coffee in hard-to-reach places. It is also helpful for forest planning, particularly through agroforestry schemes.

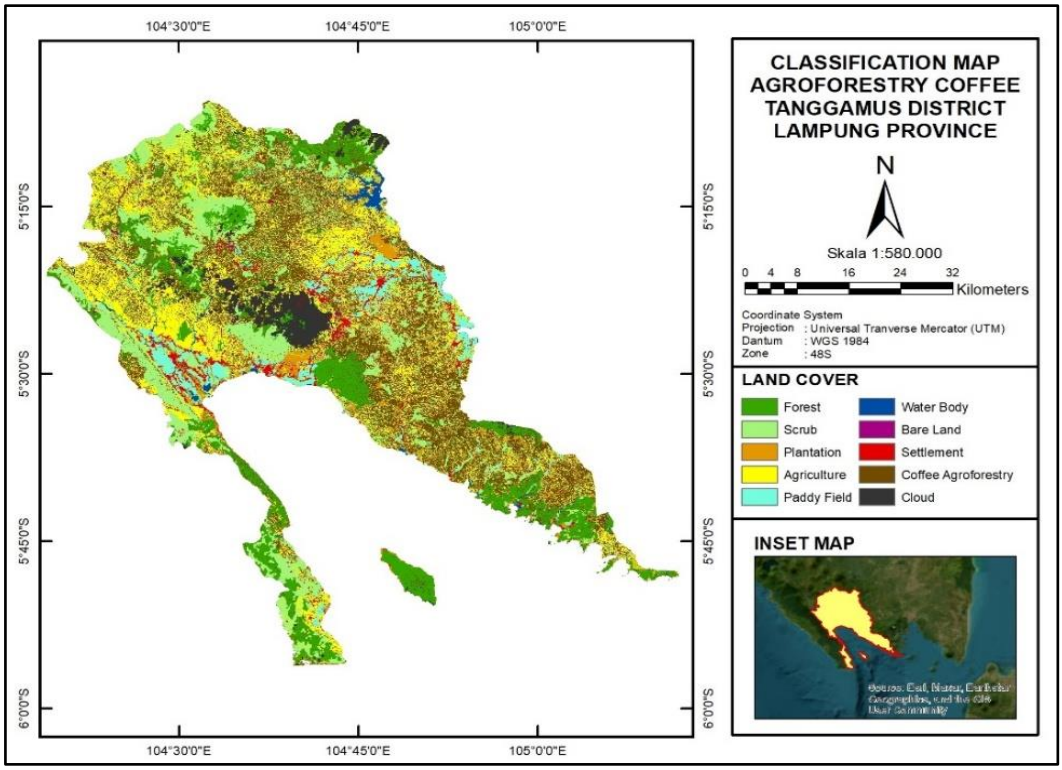


Figure 9. Spatial distribution classification map

4. CONCLUSION

This research successfully developed a coffee agroforestry classification algorithm based on ML decision trees by integrating spectral variables from SPOT-7 imagery (NDVI, VDVI, and NRGI) and biophysical variables (elevation and visual land cover). The study demonstrates that a decision-tree-based approach effectively classifies coffee agroforestry. The model with these integrated variables improved overall accuracy from 60.4% to 94.0% and kappa accuracy from 55.4% to 93.3% compared to the spectral-only model. This model also achieved user accuracy (precision) of 83.6% and producer accuracy (recall) of 89.5%. The integration of these variables proved effective in improving model accuracy while maintaining good interpretability and avoiding overfitting, thereby allowing for a clearer understanding of the determinants of coffee agroforestry systems. Slope and proximity to roads, settlements, and rivers were identified as sources of overfitting and were eliminated from the model to produce a classification output that was more logical, stable, and representative of real-world conditions. This algorithm was developed using information gain criteria without pruning or pre-pruning, with a parameter configuration of maximum depth 35, minimum leaf size 60, minimum split size 48, and a pre-pruning alternative of 60.

The resulting model can support more effective, accurate, and operational mapping of coffee agroforestry systems, particularly in the Robusta coffee production center. By integrating spectral variables from SPOT-7 imagery with elevation and visual land-cover variables, this model provides a relatively simple, readily usable, and high-performance approach, making it easy for practitioners, local governments, and research institutions to adopt without requiring complex computing infrastructure. In the context of field implementation, this model can serve as a tool for rapid, consistent spatial mapping of coffee agroforestry, especially in areas with limited data or resources. More broadly, applying this model strengthens agroforestry monitoring systems using remote sensing and ML, thereby supporting sustainable land management, environmental conservation, and increased coffee productivity. Furthermore, the model's interpretability allows policymakers to understand the key factors influencing coffee agroforestry, enabling them to formulate evidence-based policies.

This study has several limitations that require consideration when interpreting the results: (i) the model was developed using a single data source (unitemporal) with limited non-spectral variables in Tanggamus Regency, Lampung (Indonesia); (ii) the approach used focused on a single decision tree model; therefore, despite its interpretability, this model is relatively sensitive to data variation and potentially overfitting under certain conditions; (iii) this study is still based on a single observation point (unitemporal) and therefore unable to capture temporal dynamics; and (iv) model validation is still limited to a specific study area (Tanggamus Regency only). Based on these limitations, several future research directions can be developed: (i) integration of multi-sensor data (optical and RADAR imagery, or light detection and ranging (LIDAR)) with additional environmental variables (e.g., rainfall, temperature, and soil properties); (ii) development and comparison with ensemble learning approaches; (iii) integrating temporal analysis to monitor changes in coffee agroforestry over time, to support a more dynamic monitoring system; and (iv) testing the transferability of the model to other regions with different ecological and social conditions is important to ensure that the developed model is robust and widely applicable.

FUNDING INFORMATION

The authors acknowledge financial support from the research funds of the Laboratory of Remote Sensing and GIS, Department of Forest Management, Faculty of Forestry and Environment, IPB University, for field data collection in Tanggamus Regency, while no financial support was received for the publication of this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
I Made Khrisna Yoga Devandra	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
I Nengah Surati Jaya	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Tatang Tiryana	✓		✓	✓			✓			✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests or personal conflicts of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the authors upon reasonable request.

REFERENCES

[1] V. Cessa-Reyes, O. Ruiz-Rosado, and L. Alcudia-Armida, "The Coffee Agroforestry System in Mexico," *Agro Productividad*, vol. 13, no. 11, Dec. 2020, doi: 10.32854/agrop.v13i11.1811.

[2] M. O. Mardiani, I. A. Kusumawati, E. Purnamasari, C. Prayogo, M. Van Noordwijk, And K. Hairiah, "Local ecological knowledge of coffee agroforestry farmers on earthworms and their relation to soil quality in East Java (Indonesia)," *Biodiversitas*, vol. 23, no. 7, Jun. 2022, doi: 10.13057/biodiv/d230705.

[3] N. A. Ulya *et al*., "Coffee agroforestry as an alternative to the implementation of green economy practices in Indonesia: A systematic review," *AIMS Agriculture and Food*, vol. 8, no. 3, pp. 762–788, 2023, doi: 10.3934/agrfood.2023041.

[4] A. Adhiguna, I. N. S. Jaya, and N. Puspaningsih, "Developing a Decision Tree Algorithm for Detecting Agroforestry and Monoculture Coffee Plantations Using Landsat 8 Imagery: A Case Study in Bandung Regency, Indonesia," *Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan*, vol. 15, no. 6, pp. 1009–1033, Nov. 2025, doi: 10.29244/jpsl.15.6.1009.

[5] H. Ouchra, A. Belangour, and A. Erraissi, "Supervised Machine Learning Algorithms for Land Cover Classification in Casablanca, Morocco," *Ingenierie des Systemes d'Information*, vol. 29, no. 1, pp. 377–387, Feb. 2024, doi: 10.18280/isi.290137.

[6] C. Zhang, Y. Liu, and N. Tie, "Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron," *Forests*, vol. 14, no. 2, p. 254, Jan. 2023, doi: 10.3390/f14020254.

[7] T. V. Tran, R. Reef, and X. Zhu, "A Review of Spectral Indices for Mangrove Remote Sensing," *Remote Sensing*, vol. 14, no. 19, p. 4868, Sep. 2022, doi: 10.3390/rs14194868.

[8] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Overfitting, Model Tuning, and Evaluation of Prediction Performance," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.

[9] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.

[10] S. Uddin and H. Lu, "Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data," *PLoS ONE*, vol. 19, no. 4 April, p. e0301541, Apr. 2024, doi: 10.1371/journal.pone.0301541.

[11] A. B. Baloloy *et al*., "Estimation of mangrove forest aboveground biomass using multispectral bands, vegetation indices and biophysical variables derived from optical satellite imageries: rapideye, planetscope and sentinel-2," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4–3, pp. 29–36, Apr. 2018, doi: 10.5194/isprs-annals-iv-3-29-2018.

[12] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing Overfitting in Deep Networks by Decorrelating Representations," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2015, doi: 10.48550/ARXIV.1511.06068.

[13] X. Li, Y. Chen, S. Jiang, C. Wang, S. Weng, and D. Rao, "The Importance of Adding Short-Wave Infrared Bands for Forest Disturbance Monitoring in the Subtropical Region," *Sustainability*, vol. 14, no. 16, p. 10312, Aug. 2022, doi: 10.3390/su141610312.

[14] A. B. Imran *et al*., "Integration of Sentinel-2 Derived Spectral Indices and In-situ Forest Inventory to Predict Forest Biomass," *Pakistan Journal of Scientific & Industrial Research Series A: Physical Sciences*, vol. 64, no. 2, pp. 119–130, Jul. 2021, doi: 10.52763/PJSIR.PHYS.SCI.64.2.2021.119.130.

[15] A. N. Anuar, I. Jusoh, and A. Suhaili, "Vegetation Indices of Acacia Mangium Using Landsat 8 Operational Land Imager," *Journal of Sustainability Science and Management*, vol. 17, no. 4, pp. 57–69, Apr. 2022, doi: 10.46754/jssm.2022.4.005.

[16] M. A. – Rahim and Z. – S. Yu, "An In Depth Review Paper on Numerous Image Mosaicing Approaches and Techniques," *International Journal of Engineering and Management Research*, vol. 8, no. 02, Mar. 2018, doi: 10.31033/ijemr.v8i02.11607.

[17] J. Xue and B. Su, "Significant remote sensing vegetation indices: A review of developments and applications," *Journal of Sensors*, vol. 2017, 2017, doi: 10.1155/2017/1353691.

[18] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020, doi: 10.14569/ijacsa.2020.0110277.

[19] J. Li, M. Gao, and R. D'Agostino, "Evaluating classification accuracy for modern learning approaches," *Statistics in Medicine*, vol. 38, no. 13, pp. 2477–2503, Jun. 2019, doi: 10.1002/sim.8103.

[20] N. Nuthammachot and D. Stratoulias, "Exploring Sentinel-2 Satellite Imagery-Based Vegetation Indices for Classifying Healthy and Diseased Oil Palm Trees," *Journal of Oil Palm Research*, vol. 35, no. 3, pp. 517–527, Oct. 2023, doi: 10.21894/jopr.2022.0068.

[21] S. A. Salasa and W. Maharani, "Personality Detection of Twitter Social Media Users using the Support Vector Machine Method," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 2, p. 263, Dec. 2022, doi: 10.30865/json.v4i2.5345.

[22] Y. Hu, Y. Ban, Q. Zhang, X. Zhang, J. Liu, and D. Zhuang, "Spatial-temporal pattern of GIMMS NDVI and Its dynamics in mongolian plateau," in *2008 International Workshop on Earth Observation and Remote Sensing Applications, EORSA*, IEEE, Jun. 2008, pp. 1–6. doi: 10.1109/EORSA.2008.4620310.

[23] G. A. Lujan-Moreno, P. R. Howard, O. G. Rojas, and D. C. Montgomery, "Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study," *Expert Systems with Applications*, vol. 109, pp. 195–205, Nov. 2018, doi: 10.1016/j.eswa.2018.05.024.

[24] D. Kosasih, M. Buce Saleh, and L. Budi Prasetyo, "Visual and Digital Interpretations for Land Cover Classification in Kuningan District, West Java," *Jurnal Ilmu Pertanian Indonesia*, vol. 24, no. 2, pp. 101–108, Apr. 2019, doi: 10.18343/jipi.24.2.101.

[25] A. Ayanlade, "Remote sensing vegetation dynamics analytical methods: a review of vegetation indices techniques," *Geoinformatica Polonica*, vol. 16, pp. 7–17, 2017, doi: 10.4467/21995923gp.17.001.7188.

[26] N. S. Z. Suwiji, I. N. S. Jaya, N. Puspaningsih, and T. Tiryana, "Application of Machine Learning and Change Vector Analysis for Monitoring Forest Cover Change in Lore Lindu National Park, Indonesia," *European Journal of Forest Engineering*, vol. 11, no. 2, pp. 160–173, Dec. 2025, doi: 10.33904/ejfe.1675580.

[27] C. Portillo-Quintero, J. L. Hernández-Stefanoni, G. Reyes-Palomeque, and M. R. Subedi, "Novel Approaches in Tropical Forests Mapping and Monitoring–Time for Operationalization," *Remote Sensing*, vol. 14, no. 20, p. 5068, Oct. 2022, doi: 10.3390/rs14205068.

[28] M. V. Priya *et al*., "Monitoring vegetation dynamics using multi-temporal Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) images of Tamil Nadu," *Journal of Applied and Natural Science*, vol. 15, no. 3, pp. 1170–1177, Sep. 2023, doi: 10.31018/jans.v15i3.4803.

[29] T. Adugna, W. Xu, and J. Fan, "Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images," *Remote Sensing*, vol. 14, no. 3, p. 574, Jan. 2022, doi: 10.3390/rs14030574.
[30] A. Aghajanyan, "SoftTarget regularization: An effective technique to reduce over-fitting in neural networks," in *2017 3rd IEEE International Conference on Cybernetics, CYBCONF 2017 - Proceedings*, IEEE, Jun. 2017, pp. 1–5. doi: 10.1109/CYBConf.2017.7985811.
[31] D. R. Kalathoti, "Explainable AI in High-Stakes Decision Making: Beyond Accuracy," *Scientific Journal of Artificial Intelligence and Blockchain Technologies*, vol. 2, no. 3, Jul. 2025, doi: 10.63345/sjaibt.v2.i3.103.

BIOGRAPHIES OF AUTHORS

I Made Khrisna Yoga Devandra is a Master's student at the Forest Management Science Program, Department of Forest Management, IPB University, Indonesia. His academic focus centers on developing and validating digital image analysis for remote sensing. He can be contacted at email: madekhrisna@apps.ipb.ac.id, made.khrisna231@gmail.com.

I Nengah Surati Jaya is a Permanent Professor of Remote Sensing and Forest Resources Inventory at the Department of Forest Management, Faculty of Forestry and Environment, IPB University, Indonesia. His research interests include remote sensing, GIS, machine learning, forest inventory, land cover change detection, and spatial forest planning. He holds his Bachelor's degree (S1) from IPB University, and Master's and Doctoral degrees from Niigata University, Japan. He can be contacted at email: ins-jaya@apps.ipb.ac.id.

Tatang Tiryana is an Associate Professor in the Field of Forest Inventory and Biometrics at the Department of Forest Management, Faculty of Forestry and Environment, IPB University. He received a Doctoral degree from the University of Tokyo, Japan. His expertise centers on quantitative modeling of forest resources, with a focus on forest inventory, biometrics, and management planning to support sustainable forest management. He can be contacted at email: tangtir@apps.ipb.ac.id.