BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

ARTICLE

A Comparison among Different Machine Learning Algorithms in Land Cover Classification Based on the Google Earth Engine Platform: The Case Study of Hung Yen Province, Vietnam

Le Thi Lan [1,2], ***Tran Quoc Vinh*** [1], ***Phạm Quy Giang*** [3*]

[1] *Faculty of Natural Resources and Environment, Vietnam National University of Agriculture, Trau Quy, Gia Lam, Ha Noi 100000, Vietnam*

[2] *Faculty of Land Management, Hanoi University of Natural Resources and Environment, Phu Dien, Bac Tu Liem, Ha Noi 100000, Vietnam*

[3] *Faculty of Environment, Ha Long University, Uong Bi City 200000, Quang Ninh Province, Vietnam*

ABSTRACT

Based on the Google Earth Engine cloud computing data platform, this study employed three algorithms including Support Vector Machine, Random Forest, and Classification and Regression Tree to classify the current status of land covers in Hung Yen province of Vietnam using Landsat 8 OLI satellite images, a free data source with reasonable spatial and temporal resolution. The results of the study show that all three algorithms presented good classification for five basic types of land cover including Rice land, Water bodies, Perennial vegetation, Annual vegetation, Built-up areas as their overall accuracy and Kappa coefficient were greater than 80% and 0.8, respectively. Among the three algorithms, SVM achieved the highest accuracy as its overall accuracy was 86% and the Kappa coefficient was 0.88. Land cover classification based on the SVM algorithm shows that Built-up areas cover the largest area with nearly 31,495 ha, accounting for more than 33.8% of the total natural area, followed by Rice land and Perennial vegetation which cover an area of over 30,767 ha (33%) and 15,637 ha (16.8%), respectively. Water bodies and Annual vegetation cover the smallest areas with 8,820 (9.5%)

*CORRESPONDING AUTHOR:
Pham Quy Giang, Faculty of Environment, Ha Long University, Uong Bi City, Quang Ninh Province 200000, Vietnam;
E-mail: phamquygiang@daihochalong.edu.vn

ARTICLE INFO
Received: 20 May 2024 | Revised: 20 October 2024 | Accepted: 24 October 2024 | Published Online: 19 November 2024
DOI: https://doi.org/10.30564/jees.v7i1.6652

CITATION
Lan, L.T., Vinh, T.Q., Giang, P.Q., 2024. A Comparison among Different Machine Learning Algorithms in Land Cover Classification Based on the Google Earth Engine Platform: The Case Study of Hung Yen Province, Vietnam. Journal of Environmental & Earth Sciences. 7(1): 132–139. DOI: https://doi.org/10.30564/jees.v7i1.6652

ha and 6,302 ha (6.8%), respectively. The results of this study can be used for land use management and planning as well as other natural resource and environmental management purposes in the province.
Keywords: Google Earth Engine; Land Cover; Landsat; Machine Learning Algorithm

1. Introduction

Maps of the current status of land cover are very necessary documents for carrying out land statistics and land inventory work, and at the same time, they are useful information for the management and supervision of land use implementation. The application of information technology combined with Remote Sensing technology is one of the effective, fast, and cost-effective solutions to support the creation of land use status maps. Especially, using satellite data can determine land use status in almost real time.

Some previous traditional algorithms such as Maximum Likelihood Classifier (MLC), Minimum Distance Classifier (MDC), and K-Nearest Neighbor (KNN) are often applied in commercial software such as ArcGIS, Erdas Imagine, Envi, ER Mapper, etc. [1, 2]. Currently, the strong development of the 4.0 revolution has created many advantages for building the current status map of land cover from satellite images by using machine learning algorithms, including outstanding algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Classification and Regression Tree (CART)[3–5].

Many studies in the world have used machine learning algorithms to classify land cover from satellite images, such as Hamad[6], Yuh et al.[7], Cai et al.[8], Biswas et al.[9], Mhanna et al.[10], and Mollick, Azam and Karim[11]. In Vietnam, the application of machine learning algorithms in land cover classification and mapping is also very popular. The studies by Bui and Trinh[12], Dang[13], and Nguyen et al.[14] can be mentioned as typical examples. The results from these studies have demonstrated the effectiveness of machine learning in determining the current status of land cover. However, recent studies mostly used a single method instead of a comparison among the classification methods to select the best method for the classification. This article presents the results of classifying the current status of land cover of Hung Yen province in northern Vietnam using Landsat satellite images based on the Google Earth Engine (GEE) cloud computing platform and the application of three machine learning algorithms, including Support Vector Machine (SVM), Random Forest (RF), and Classification and Regression Tree (CART).

2. Methodology

2.1. The Study Area

Hung Yen Province is located in the Red River Delta, northern Vietnam. The province covers an area of over 930 km^2, comprising 8 districts, 1 district-level town, and 1 city with a population of 1,302,000 people (in 2022), and an average population density of 1,400 people km^{-2} (ranked 4th in the country). Situated in the Red River Delta, one of the two deltas of Vietnam (the other is the Mekong River Delta), the province has flat terrain and fertile land, and used to be an agricultural province. However, being located in the northern key economic region, the center of the development triangle Ha Noi—Hai Phong—Quang Ninh, the province has been developing towards increasing services and industry. In recent years, the demand for land use from industrial economic sectors has been very large, leading to the conversion of agricultural land use purposes to industrial ones, which resulted in narrowing agricultural land area. Therefore, research on land cover is important to help managers and planners make decisions on land use effectively, economically, and sustainably. The geographic location of Hung Yen Province is presented in **Figure 1**.

2.2. Methods

2.2.1. Land Cover Classification on GEE

In this study, three machine learning algorithms including Support Vector Machine, Random Forest, and Classification And Regression Tree were applied to calculate and classify land cover using Landsat satellite images. SVM is a supervised machine learning algorithm that typically delivers good results in classification and regression. SVM divides support vectors to classify log data points to find two types of independent support vectors with the largest amplitude [15, 16]. The CART algorithm is based on a decision tree classification system and uses training samples to identify, recognize, and classify objects on satellite images. Decision trees include multi-level and multi-leaf nodes. Maximum nodes refer to the maximum number of leaves per tree, and the minimum leaf population is the minimum number of nodes created for the training set only. To build a suitable tree, enough nodes and branches must be created. The maximum node value is unlimited if it is not specified. Meanwhile, RF is an integrated learning algorithm that can integrate multiple decision trees and then form a forest. The algorithm combines random features to create a tree. The bagging method is used to generate training samples and each selected feature is drawn randomly by replacing the size of the initial training set. Then, the final prediction result is obtained by combining multiple decision trees.

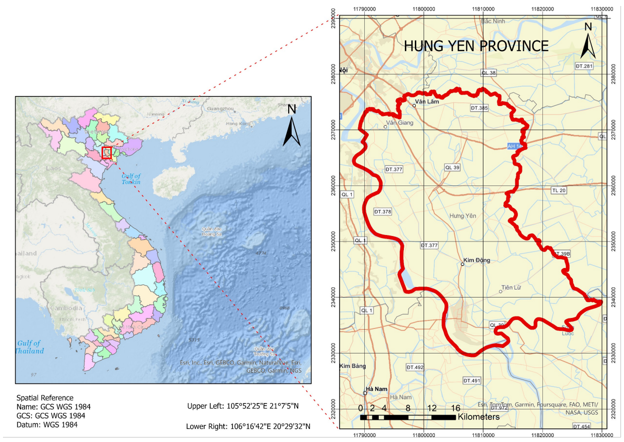

Figure 1. The geographic location of Hung Yen Province.

The SVM, RF, and CART are among the most widely applied and proven algorithms, which have demonstrated outstanding effectiveness in classifying land cover from satellite image data [17]. The study area was selected for a classification test with 6 basic land layers including: 1) Rice land, 2) Water bodies, 3) Perennial vegetation, 4) Annual vegetation, 5) Built-up areas, and 6) transportation. The land covers selected for classification are current at the time of image acquisition. Characteristics of the selected Bare land are areas such as soil, sand, mudflats, areas being leveled, and newly leveled bare land areas preparing for construction. Water bodies include ponds, rivers, streams, canals, and aquacultural land. Perennial vegetation includes areas with fruit trees, shade trees, timber trees, mixed gardens, ornamental flowers, and gardens interspersed in residential areas. Annual vegetation includes areas growing rice, vegetables, grasslands, medicinal plants, etc. Built-up areas include areas of houses, apartments, other non-residential constructions, industrial parks, industrial clusters, factories, warehouses, cemeteries, and temples... Transportation land cover includes main traffic routes such as national highways, provincial roads, district roads, inter-commune roads, and intra-field roads... The total number of sampling points is 640 points for 6 land cover types of the entire study area. The implementation of land cover classification was conducted on the GEE cloud computing data platform with the JavaScript programming language (**Figure 2**).

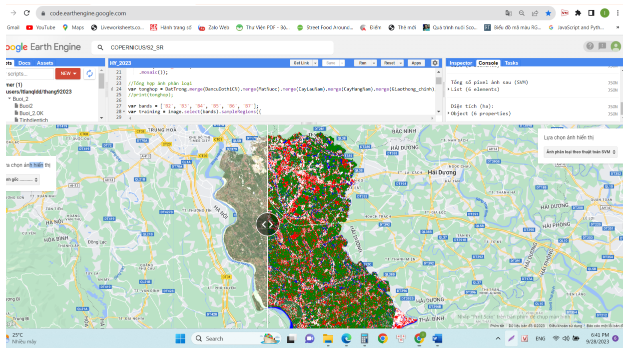

Figure 2. GEE code writing and command execution interface.

The process was as follows: first, obtaining Landsat satellite image data of the study area; then filtering clouds to ensure the best quality of data; conducting training for the machine learning program and recording information about samples for each type of land cover; classifying land cover according to trained samples; recording results and evaluating classification accuracy. An overview of the land cover classification process is illustrated in **Figure 3**.

2.2.2. Accuracy Assessment

The accuracy assessment of land cover classification in this study was done using the Confusion Matrix method. A confusion matrix (or error matrix) is a table that shows the correspondence between the classification result and reference data. This method is commonly used in Remote Sensing as the quantitative method to represent the difference between the actual and predicted classifications. Basic statistics for the confusion matrix include an assessment of overall accuracy (Overall Accuracy—OA) and the Kappa coefficient.

The overall classification accuracy is calculated by dividing the total number of correctly classified values by the total number of values.

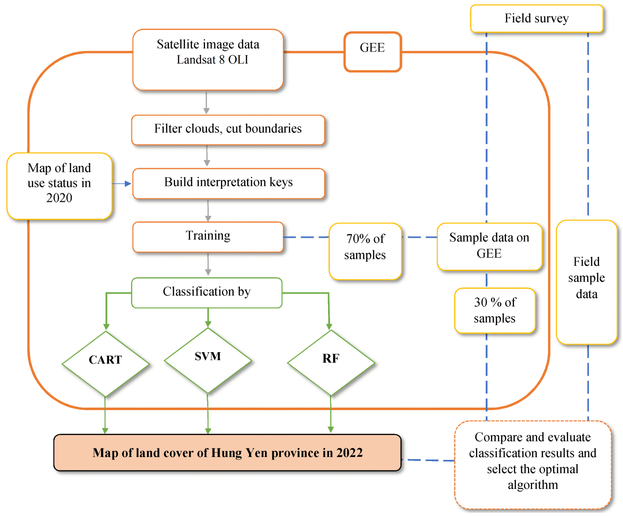

Figure 3. Land cover classification process on GEE using CART, SVM, and RF algorithms.

The Kappa coefficient measures the agreement between classification and truth values. The Kappa value varies from 0 to 1. A kappa value of 0 represents no agreement, while the closer the OA value is to 1, the more reliable it is; the Kappa coefficient with a value from 0.4 to 0.6 is considered as "average", from 0.6 to 0.8 is considered as "good", and more than 0.8 to 1.0 is considered as "very good"[16].

3. Results and Discussion

The results of land cover classification according to the CART, RF, and SVM algorithms are illustrated in **Figure 4**. There are 5 basic cover layers: Rice land, Water bodies, Perennial vegetation, Annual vegetation, and Built-up areas.

The classification results are shown in terms of the area of each cover type, and the error matrix for accuracy assessment (**Table 1**).

In this study, 640 ground points were used as training samples for classification and reference points for accuracy assessment, of which 70% of the points (448 points) were used as training samples and 30% of the points (192 points) were used for accuracy assessment. Training samples are primarily collected on a per-pixel basis to reduce redundancy and spatial autocorrelation. The points were selected through image interpretation with intensive field visits over the study area. According to Gong and Howarth[18] and Foody et al.[19], more training samples tend to be more representative of the class population so the more training samples, the better. However, a small number of training samples is obviously attractive for logistic reasons. Many previous studies recommended that for classifiers that require few parameters to be estimated like the maximum likelihood when applied to a handful number of bands, the number of training samples for each class should be from 10 to 30 times the number of bands[20–22]. For many classification algorithms, no previous study has reported an optimal number of training samples. There are also other studies that have proven that the number of classification samples is not necessarily too large, the number of samples only needs to be 2 to 4 times the number of image bands used in classification. Even with that number of samples, the overall accuracy and Kappa coefficient can reach as high as over 90%[23, 24].

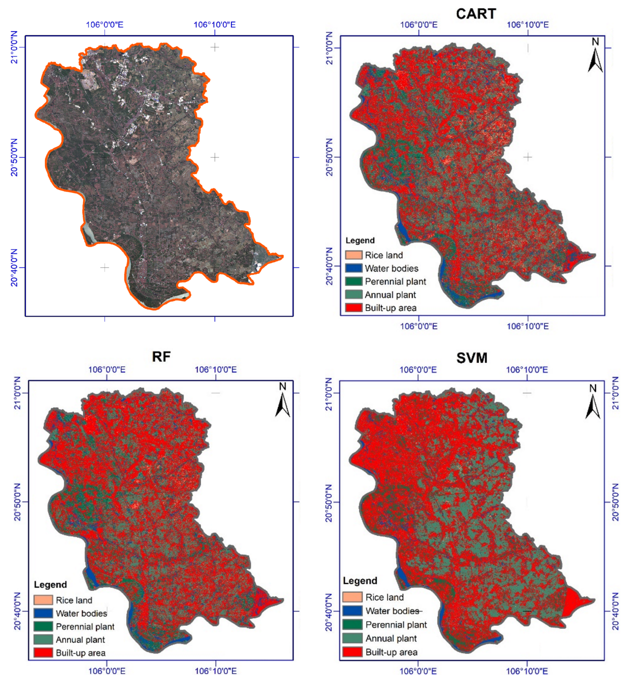

Figure 4. Original image and map of land cover types established by CART, RF, and SVM algorithms.

For this study, the study area is not too large, the Water bodies, Annual vegetation, and Perennial vegetation have many uniform pixel values and 6 bands of Landsat were used for classification, so the number of samples used for each class was 91 samples. Only Water bodies areas have a lesser number of samples (84 samples) because it has more characteristic pixel values than the others. On average, the total number of samples was equivalent to more than 75 times the number of image bands used. The classification by CART, RF, and SVM algorithms on GEE got an overall accuracy

Table 1. Confusion matrix and classification accuracy by SVM.

Land Cover	Rice Land	Annual Vegetation	Perennial Vegetation	Built-Up Areas	Water Bodies	Total	User's Accuracy
Rice land	34	2	1	2	0	39	87.2%
Annual vegetation	2	34	1	1	1	39	87.18%
Perennial vegetation	1	1	35	2	1	40	87.5%
Built-up areas	1	1	1	33	1	37	89.2%
Water bodies	1	1	1	1	33	37	89.2%
Total	39	39	39	39	36	192	
Producer's accuracy	87.2%	87.2%	89.7%	84.6%	91.7%		
Overall accuracy (OA): 88.02%						Kappa: 0.86	

(OA) of 81%, 83%, and 88%, respectively, and the Kappa coefficient of 0.80, 0.82, and 0.86, respectively. According to Visa et al.[16], the accuracy of all three algorithms can be considered as "very good", but among the three algorithms, SVM got the most accurate classification with an overall accuracy of 88% and a Kappa coefficient of 0.86. In fact, SVM is known for its effectiveness in handling high-dimensional data and complex decision boundaries. It works well with small to medium-sized datasets and is particularly suitable for binary and multiclass classification tasks. Therefore, it is understandable that in this case study, SVM achieved the most accurate result in comparison with the other two algorithms. The high accuracy of all three algorithms in this study can also be explained by the quality of the images. Training samples and image quality have a certain influence on classification accuracy. In this study, the image obtained has very low cloud coverage (less than 2%) and clear physical information. A confusion matrix for the calculation of classification accuracies by the SVM algorithm is shown in **Table 1**.

After evaluating using the confusion matrix on the GEE platform to ensure accuracy, the study has also conducted field sampling points for additional evaluation in 10 districts with the support of handheld GPS combined with land use status maps to verify land use types on classified images. Satellite imagery classification algorithms assign pixels to different land cover classes based on spectral signatures. Field sampling provides the means to verify whether the classifications are accurate. Field sampling helps identify errors or discrepancies between the classified land cover types and the actual land cover on the ground[25]. These errors could arise due to factors such as spectral confusion, misclassification, or changes in land cover over time. At the sampling sites, we collected a total of 60 land use samples, which do not overlap with the samples on GEE. The verification result shows that 4 points were misclassified as other land use types, while the remaining 56 points (accounting for 93.3% of the total number of points) were consistent with the results of classified land use types from Landsat 8 satellite images. Typically, accuracy values above 80% are considered acceptable. In this case, the accuracy of 93.3% shows very good performance of the classification by GEE.

In short, from the accuracy assessment based on the confusion matrix on the GEE platform and field verification, it can be concluded that land use classification was accurate. A high classification accuracy ensures that the classification results are reliable for land use management when necessary.

The results of land cover classification by CART, RF, and SVM algorithms on the GEE platform with a comparison to the land use data reported by the local government are shown in **Table 2** and summarized in **Figure 5**. According to the land statistics of the Hung Yen provincial government, as of December 31, 2022, the total natural land area of the entire province was 93,019.80 hectares. This figure is 0.1 ha larger than the total land area classified by the three algorithms on GEE, which is 93,019.70 ha. This difference occurred during the image-cutting process using the province boundary as a mask layer. In addition, there were null pixels in the images, a common problem due to satellite sensor malfunction and poor atmospheric conditions[26]. Null pixels can create spatial discontinuities in the image and the classification algorithms may struggle to accurately classify areas with missing data, leading to misclassifications. In general, in comparison with the government-reported data, the SVM algorithm shows the smallest difference with Annual vegetation land cover being the largest difference (93.99 ha, which is equal to 1.51% of the area of Annual vegetation reported by the local government). For the CART algorithm, the largest difference occurs with Rice land (1,151.59 ha, which is equal to 3.73%), while for the RF algorithm, the largest

Table 2. Summary of land cover classification by the three algorithms.

Land Cover Types	Government Data (ha)	SVM			CART			RF		
		Area (ha)	Difference (ha)	%	Area (ha)	Difference (ha)	%	Area (ha)	Difference (ha)	%
Rice land	30,845.60	30,767.35	–78.25	–0.25	29,694.01	–1151.59	–3.73	31,693.25	847.65	2.75
Annual vegetation	6,207.90	6,301.89	93.99	1.51	6,045.66	–162.24	–2.61	6,243.02	35.12	0.57
Perennial vegetation	15,701.20	15,637.30	–63.9	–0.41	16,231.91	530.71	3.38	15,950.05	248.85	1.58
Built-up areas	31,456.80	31,494.08	37.28	0.12	32,052.66	595.86	1.89	30,490.68	–966.12	–3.07
Water bodies	8,808.30	8,819.08	10.78	0.12	8,995.46	187.16	2.12	8,642.7	–165.6	–1.88
Total	93,019.80	93,019.70	–0.10		93,019.70	–0.10		93,019.70	–0.10	

difference occurs with Built-up areas (966.12 ha, which is equal to 3.07%).

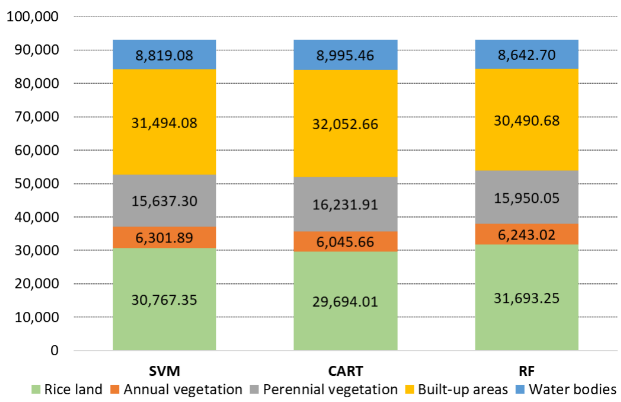

Figure 5. Area and proportion of land cover types in Hung Yen province.

The classification result from the SWM algorithm, which is the most accurate algorithm in this study, shows that the Water bodies in the study area cover an area of nearly 8,820 ha, accounting for nearly 9.5% of the total natural area, of which the Red River part in the province accounts for about 80%–90% of Water bodies. Annual vegetation and Perennial vegetation cover an area of nearly 6,302 ha (6.8%) and over 15,637 ha (16.8%), respectively. In Hung Yen Province, annual vegetation mainly includes vegetables, maize, flowers, and ornamental plants, whereas Perennial vegetation mainly includes fruit trees such as longan, grapefruit, guava, and lychee. Built-up areas cover the largest area with nearly 31,495 ha, accounting for more than 33.8%, followed by Rice land, which covers an area of over 30,767 ha, accounting for more than 33% of the total natural area. It is understandable that Built-up areas cover the largest among all the land cover types because Hung Yen is one of the most populated provinces with a population density of 1,350 people km^{-2} (ranking 4th in the country after Ho Chi Minh City, Ha Noi, and Bac Ninh) and a very fast urbanization process[27]. The large Rice land area indicates that agriculture remains an important sector of this province. The result of the above land cover classification is a useful source of data that can be referenced to help land management and land use planning. Governments, urban planners, and environmental management agencies could use these data to make informed decisions regarding zoning regulations, resource allocation, and sustainable development practices.

4. Conclusions

Exploiting and analyzing online satellite image data for land management is becoming increasingly effective, fast, and cost-effective. In addition, applying machine learning algorithms in classifying land cover from satellite images also actively contributes to the digital transformation and industrial revolution 4.0 in general. With support from the Google Earth Engine platform integrated with the SVM, CART, and RF algorithms to interpret Landsat 8 satellite images for Hung Yen province, the study has produced an accurate and reliable classification of land cover which includes six land cover types: Bare land, Water bodies, Annual vegetation, Perennial vegetation, Built-up areas, and Transportation. In particular, the SVM algorithm got a great performance as its overall accuracy reached 86% and its Kappa coefficient was 0.88. The classification results are detailed and timely and can therefore be used for supervising land use planning, monitoring, and forecasting land use changes.

Author Contributions

L.T.L. contributes to conceptualization, data collection, formal analysis, writing original draft; P.Q.G. contributes to conceptualization, writing original draft, formal analysis, methodology, visualization, review and editing; T.Q.V. contributes to conceptualization, methodology, review and editing. The authors contributed equally to this study.

Funding

This work received no external funding.

Data Availability Statement

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

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Fischer, M., Arthur, G., 2010. Handbook of applied spatial analysis: Software tools, methods and applications. Springer: Berlin, Germany. pp. 125–134.

[2] Rajan, M., Navin, S., Loganathan, A., et al., 2020. Survey on Land Use/Land Cover (LU/LC) change analysis in remote sensing and GIS environment: Techniques and Challenges. Environmental Science and Pollution Research. 27(24), 29900−29926.

[3] Sellami, E.L.M., Rhinane, H., 2022. A new approach for mapping land use/ land cover using Google Earth Engine: A comparison of composition images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 48, 343.

[4] Ouchra, H., Belangour, A., Erraissi, A., 2023. Machine learning algorithms for satellite image classification using Google Earth Engine and Landsat satellite data: Morocco case study. IEEE Access. 11, 71127−71142.

[5] Ramachandra, T.V., Mondal, T., Setturu, B., 2023. Relative performance evaluation of machine learning algorithms for land use classification using multispectral moderate resolution data. SN Applied Sciences. 5, 274.

[6] Hamad, R., 2020. An assessment of artificial neural networks support vector machines and decision trees for land cover classification using sentinel-2A. Data Science. 8, 459.

[7] Yuh, Y.G., Tracz, W., Matthews, H.D., et al., 2023. Application of machine learning approaches for land cover monitoring in northern Cameroon. Ecological Informatics. 74, 101955.

[8] Cai, Y., Guan, K., Peng, J., et al., 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sensing of Environment. 210, 35.

[9] Biswas, J., Jobaer, M.A., Haque, S.F., et al., 2023. Mapping and monitoring land use land cover dynamics employing Google Earth Engine and machine learning algorithms on Chattogram, Bangladesh. Heliyon. 9(11), e21245.

[10] Mhanna, S., Halloran, L.J.S., Zwahlen, F., et al., 2023. Using machine learning and remote sensing to track land use/land cover changes due to armed conflict. Science of The Total Environment. 898, 165600.

[11] Mollick, T., Azam, M.G., Karim, S., 2023. Geospatial-based machine learning techniques for land use and land cover mapping using a high-resolution unmanned aerial vehicle image. Remote Sensing Applications: Society and Environment. 9, 100859.

[12] Bui, T.H.T., Trinh, T.T., 2020. Classification of surface cover extraction objects in the Non nuoc Cao Bang global geopark area based on cloud computing platform. Journal of Natural Resources and Environmental Sciences. 31, 65.

[13] Dang, T.T., 2021. Online exploitation of satellite image databases, comparison of machine learning algorithms on land cover classification on Google Earth Engine platform. In Proceedings of the National Conference on Solutions to Connect and Share Database Systems for Training and Management in the Field of Natural Resources and Environment; 21 December 2021; Hanoi University of Natural Resources and Environment: Ha Noi, Vietnam. p. 192.

[14] Nguyen, T.T., Phan, V.T., Nguyen, V.Q., et al., 2022. Using Sentinel 2 satellite images and machine learning algorithm to establish the current status map of forests in Bu Dang District, Binh Phuoc Province. Can Tho University Journal of Science. 58(6B), 150.

[15] Abdi, A.M., 2020. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GIScience Remote Sensing. 57, 1.

[16] Visa, S., Ramsay, B., Ralescu, A.L., et al., 2011. Confusion matrix-based feature selection. Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference; Cincinnati, OH, USA; 16–17 April 2011. p. 120.

[17] Zhewen, Z., Fakhrul, I., Liaqat, A.W., et al., 2024. Comparison of three machine learning algorithms using Google Earth Engine for Land Use Land Cover classification. Rangeland Ecology & Management. 92, 129.

[18] Gong, P., Howarth, P.J., 1990. An assessment of some factors influencing multispectral land-cover classification. Photogrammetric Engineering & Remote Sensing. 56, 597.

[19] Foody, G.M., Mathur, A., Sanchez-Hernandez, C., et al., 2006. Training set size requirements for the classification of a specific class. Remote Sensing of the Environment. 104, 1–14.

[20] Van Niel, T., Mcvicar, T., Datt, B., 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-

temporal classification. Remote Sensing of the Environment. 98, 468.

[21] Zhang, H., He, J., Chen, S., et al., 2023. Comparing three methods of selecting training samples in supervised classification of multispectral remote sensing images. Sensors. 23(20), 8530.

[22] Phinzi, K., Ngetar, N.S., Pham, Q.B., et al., 2023. Understanding the role of training sample size in the uncertainty of high-resolution LULC mapping using random forest. Earth Science Informatics. 16, 3667.

[23] Li, C., Wang, J., Wang, L., et al., 2014. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. Remote Sensing. 6(2), 964.

[24] May, F.P., Santos, R.A., Toledo, G.R., et al., 2022. Mapping of land cover with optical images, supervised algorithms, and Google Earth Engine. Sensors. 22(13), 4729.

[25] Kamusoko, C., 2022. Land cover classification accuracy assessment. In: Kamusoko, C. (Ed.). Optical and SAR Remote Sensing of Urban Areas: A Practical Guide (Springer Geography). Springer: Singapore. p. 105.

[26] Chen, M., Sun, Z., Newell, B.H., et al., 2020. Missing pixel reconstruction on Landsat 8 Analysis Ready Data Land Surface Temperature image patches using Source-Augmented Partial Convolution. Remote Sensing. 12(19), 3143.

[27] General Statistics Office of Vietnam, 2023. Statistical yearbook of 2022. Statistical Publishing House: Hanoi, Vietnam. pp. 41–100.