

The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping

Hai Ly NGO^{1,2,3,4}, Huu Duy NGUYEN^{3*}, Peio LOUBIERE², Truong Van TRAN³, Gheorghe ȘERBAN⁵, Martina ZELENKOVA⁶, Petre BREȚCAN⁷ and Dominique LAFFLY¹

Authors' affiliations and addresses:

¹ Department of Geography, Planning and Environment, University of Jean Jaures, Toulouse
e-mail: lngo@ewmi-odi.org
e-mail: dominique.laffly@univ-tlse2.fr

² CY Techs, Pau, France
e-mail: peio.loubiere@cyu.fr

³ Faculty of Geography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Vietnam
e-mail: nguyenhuuduy@hus.edu.vn;
e-mail: tranvantruong@hus.edu.vn

⁴ Open development Mekong

⁵ Faculty of Geography, Babes-Bolyai University, Cluj-Napoca, Romania
e-mail: gh.serban@icloud.com

⁶ Department of Environmental Engineering, Faculty of Civil Engineering, Technical University of Kosice, Slovakia
e-mail: martina.zelenakova@tuke.sk

⁷ Department of Geography, Faculty of Humanities, Valahia University of Târgoviste, 130105 Dambovita, Romania
e-mail: petrebretcan@yahoo.com

*Correspondence:

Huu Duy Nguyen, Faculty of Geography, VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan district, Hanoi, Vietnam
e-mail: nguyenhuuduy@hus.edu.vn

How to cite this article:

Ngo, H.L., Nguyen, H.D., Loubiere, P., Tran, T.V., Șerban, G., Zelenakova, M., Brețcan, P. and Laffly, D. (2022). The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping. *Acta Montanistica Slovaca*, Volume 27 (2), 342-359.

DOI:

<https://doi.org/10.46544/AMS.v27i2.05>

Abstract

Monitoring Land-use/land-cover (LULC) changes are a significant challenge for sustainable spatial planning, particularly in response to transformation and degenerative landscape processes. These disturbances lead to the vulnerability of inhabitants and habitat and climate changes and socio-economic development in the region. Several studies have proposed different methods and techniques to monitor the spatial and temporal changes of LULC. Machine learning is a more popular method. However, the problem of data imbalance is a significant challenge, and the classification results tend to bias the majority classes for unbalanced data. Therefore, this study's objective is to develop a state-of-the-art technique to reduce the problem of data imbalance in LULC classification in Vietnam based on machine learning and SMOTE (Synthesizing Minority Oversampling Technology) with Edited Nearest Neighbor (ENN). Various statistical indices, including Kappa and Accuracy, have been used to assess the performance for the classification of Land-use/cover. The results indicate that integrating oversampling and under-sampling with SMOTE ENN gave better overall accuracy and generalization. We also find that the expected proportion of chance agreement after oversampling is higher than before (Kappa score before and after oversampling is 0.905244 and 0.974379, respectively). This study provides an effective method to monitor spatial and temporal land cover change in Vietnam; it plays a role as a framework for other relevant research related to land cover change, which can support planning and sustainable management of the territory.

Keywords

LULC mapping, SMOTE ENN, Machine learning, imbalance data



© 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

In the context of population growth, environmental changes and feedback, and increasing global interconnectedness due to globalization processes, understanding the diverse and complex interactions of socio-economy land systems are still significant challenges to global scientific communities (Mohamed et al., 2020). Change in Land Use and Land Cover (LULC) can cause direct or indirect effects on the natural environment, biodiversity, quality of water, land and air resources, ecosystem functions, and services, making essential effects on the climate system (Lambin et al., 2000; Prakasam 2010; Petrişor et al., 2020; Aouissi et al., 2021). Currently, LULC statistics are the prerequisites for decision-makers to build the strategies and policies that influence the importance of societies and their economies (Reidsma et al., 2011). For these reasons, the study of LULC changes plays a vital role in debates on sustainable development.

Typically, land-use/cover changes are visible in remotely-sensed data or by data derived from secondary statistics like agricultural census data. Under certain circumstances, land-use/land-cover changes can infer in remotely-sensed data (Lambin and Geist, 2008). Recently, the satellite-based observations of the Earth have provided a valuable data source for monitoring land-use/cover changes at spatially and temporally scale at pixel resolution. In 2008, the USGS adopted a free and open Landsat data policy (Turner et al., 2015), and in 2014, the European Union also proposed a policy of full, open and free access to data from the Copernicus program. There are many remote sensing datasets in this data pool, such as GLS (Global Land Survey), Landsat collection, and Sentinel-2 collection. GLS is an ortho-rectified and cloudless data source at a resolution of 30 meters at times, representing 1975, 1990, 2000, 2005 and 2010. The observation infrastructures acquire data in the nominal year of each period (Franks et al., Gutman et al., 2008). Sentinel-2 collection is still in infancy, and the period of long time series of this one is limited more than the Landsat collection. The studies on monitoring land-use/cover changes considered Landsat data standard data (Lambin and Geist, 2008).

The traditional LULC classification method is based on on sample identification (Brennan and Webster, 2006, Platt and Rapoza, 2008, Mohamed et al., 2020). However, extracting many features in traditional supervised classification is time-consuming and requires a deep understanding of natural geography, spatial features and the experience of the operator. In addition, the supervised classification is done by hand in most cases, which leads to insufficient sample size and low generalization ability. The limited samples also influence the accuracy and reliability of the classification results.

Recently, the world scientific community has been interested in using machine learning algorithms for classifying land-use/cover, such as support vector machine (SVM) (Huang et al., 2002; Shao and Lunetta, 2012), Random forest (RF) (Rodriguez-Galiano et al., 2012; Thanh Noi and Kappas 2018), Artificial Neural Network (ANN) (Song et al., 2012; Kadavi and Lee 2018), Extreme Gradient Boosting (XGB) (Georganos et al., 2018), cat boost (CB). However, these methods have been proven effective for LULC classification. However, the data imbalance problem is still a massive challenge for the scientific community. An unbalanced dataset is the one in which there are several examples in one class, known as the minority class, and hundreds, thousands, or millions of examples in the other ones, known as the primary class (Galar et al., 2011; Yijing et al., 2016). The solution to dealing with the imbalanced data problem has three approaches: data level approaches, algorithm level, and cost-sensitive methods (Galar et al., 2011; López et al., 2013). The data level approach is a preprocessing solution before building a classification model (Vluymans, 2019). Three resampling categories are under-sampling methods, over-sampling methods, and hybrid methods. Under-sampling is a technique that removes examples from the original dataset to decrease the imbalance ratio. Oversampling is the technique that takes the opposite approach to under-sampling. Instead of reducing the majority class, the number of instances in the minority class increases. The new ones will be added to the training dataset with orientation, such as duplicating original instances or interpolating elements in the minority classes. Hybrid methods are the ones combined between oversampling and under-sampling (Vluymans 2019). In the first case, the under-sampling methods can remove the potential or valuable data essential for constructing a predictive classification model. With random oversampling, the likelihood of overfitting is very high when it copies the elements precisely in the dataset (Fernández et al., 2018). One popular oversampling method known is SMOTE (Synthetic Minority Over-sampling Technique). (Batista et al., 2004) proved that the oversampling methods gave them better results than under-sampling methods in considering the area of the ROC (Receiver Operating Characteristic) curve (AUC). The authors highlighted that SMOTE + ENN also provided excellent results for a small dataset of the minority. The technique could eliminate the noise examples in the dataset. SMOTE + ENN is a collaboration between SMOTE and ENN (Edited Nearest Neighbor Rule). It tends to give a production of data cleaning in depth. That is the reason we selected the SMOTE ENN in our study.

This study aims to develop a method to monitor spatial and temporal LULC changes in Vietnam based on machine learning, SMOTE ENN, and Remote Sensing. The initial hypothesis tested in this study is that SMOTE ENN successfully reduces the data imbalance problem to improve the performance of the classification model. This study can be considered a framework to minimize field costs and processing techniques to obtain the best land use and land cover classification results with unbalanced near real-time remote sensing data at the national

and regional levels. The results of this study can support decision-makers in developing strategies and policies for the planning and sustainable management of the territory.

Material and Methods

Study areas

The study area includes parts of Quang Tri province, Thua Thien Hue province, Da Nang city, and parts of Quang Nam province in the Central region of Vietnam shown in Figure 1. The topography of the study area consists of coastal plains and inland mountainous areas and narrow plains in valleys. In the south, the mountain ranges face the sea east-west direction. The rivers in the region divide the mountainous terrain very strongly, so every few tens of kilometres have an estuary flowing into the sea. The Truong Son Mountain range passes this area, and there is the Bach Ma peak of 1444m.

The study area is located mainly in the northern climate zone of Vietnam. There is a small part of the south that belongs to the southern climate zone of Vietnam. First, the Bach Ma mountain range and Hai Van pass are the precise boundaries of climate zoning between the two North and South Vietnam regions in the study area. The climate in the study area in Thua Thien Hue province north of Hai Van pass is transitional from subtropical to tropical monsoon, without a distinct winter and dry season. In the south, it is located within the boundary of Da Nang city and is divided into two seasons: the rainy season and the dry season. The rainy season starts from October to December.

In contrast, the dry season starts from January to September. We collect remote sensing data from April to September because, during the rainy season, the high cloud coverage affects the image acquisition process. Second, the area of wet rice cultivation reduces due to inundation. The study area is contiguous to the sea and mountains and has plain topography; it is typical for generalization. Land use types include bare land, river or water land, urban land, field land or paddy land, crops land, orchards land, and forest land.

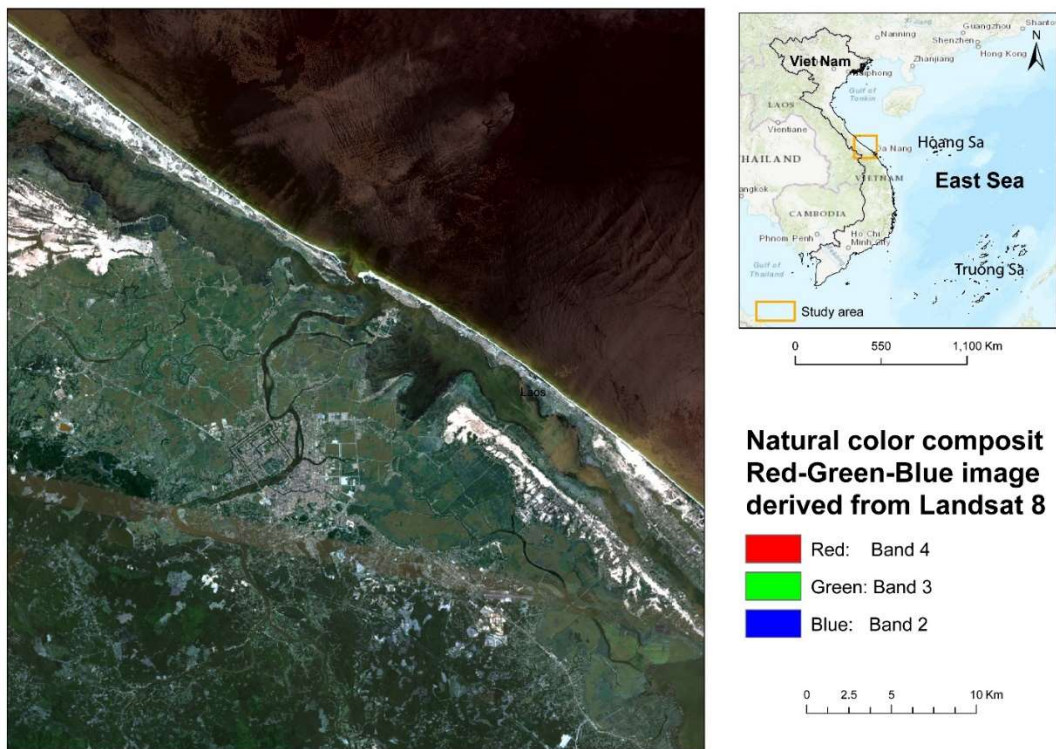


Fig. 1. Location of study areas in Vietnam

According to regulations in Vietnam, the category is divided into three primary types of land-use: Agricultural land, Non-agricultural land, and non-use land at 4 level sub-categories corresponding to the administration level: National, provincial, district, and communal levels. Besides, according to the Japan Aerospace Exploration Agency (JAXA), at the national scale, they divided into seven classes of land use/cover for classification: bare soil, water, built-up, paddy, crops, orchards, and forest.

Pre-processing of Satellite image

In this study, Landsat 8 OLI 1 bands with a wavelength from 0.4430 to 2.2010 μm including coastal aerosol (0.4430), Blue (0.4826), Green (0.5613), Red (0.6546), Near Infrared (0.8646), Shortwave Infrared 1 (1.6090),

and Short-wave Infrared 2 (2.2010) from April to September 2016 with the cloud cover less than 30 percent was used as the input data set in which we used 14 scenes from the 2016 time series of images, six natural bands (B2, B3, B4, B5, B6, B7), nine indicator bands (NDWI, EVI, GVI, EBBI, SAVI, SBI, PCA1, PCA2, PCA3), and one support band is DEM (Table 1 and Table 2). The data was uploaded to the site <https://earthexplorer.usgs.gov>. The Landsat images are available at a spatial resolution of 30 m. For preprocessing, Atmospheric correction processes and the incorporation of high spatial resolution and panchromatic band images (15m) play an essential role in the remote sensing mapping process. Here fusion technique is applied to produce higher resolution images with spectral information for multiple bands. With the development of science and technology, more and more techniques are used, such as Wavelength transformation, Brovery transformation, Principal component transformation, and High pass filtering (Bui et al., 2021). In this study, the Gram-Schmidt method in ENVI software is used to integrate images based on the modelling of Phanchromatic bands and calculate the average value of multispectral bands. This method has proven effective in optimizing sharpness and minimizing colour distortion in images (Bui et al., 2021). For the image classification task, from the six multispectral bands in the Landsat 8 OLI image, the spectral numbers include Enhanced Vegetation Index (EVI), Enhanced Built-Up and Bareness Index (EBBI), Global Vegetation Index (GVI), Normalized Difference Water Index (NDWI), Soil-Adjusted Vegetation Index (SAVI), Soil Brightness Index (SBI) is used to improve the classification of satellite images. These indicators are effective in previous studies. We also use principal component analysis to select three bands which have the most information.

Tab. 1. Formulas of some spectral indicators

Indicators	Formula	Notes
NDWI	$(GREEN - NIR)/(GREEN + NIR)$	Elimination of the soil and terrestrial features would facilitate the delineation of open water (Bhatti et al., 2014)
EVI	$2.5 * \frac{Band\ 5 - Band\ 4}{Band\ 5 + 6 * Band\ 4 - 7.5 * Band\ 2 + 1}$	Landsat Enhanced Vegetation Index (EVI) is similar to Normalized Difference Vegetation Index (NDVI) and can be used to quantify vegetation greenness. However, EVI corrects for some atmospheric conditions and canopy background noise and is more sensitive in areas with dense vegetation.
GVI	$- 0.2848*Band2 - 0.2435*Band3 - 0.54363*Band4 + 0.7243*Band5 + 0.0840*Band6 - 0.1800*Band7$	The greenness image or green vegetation index (GVI) presents high values in targets with a high density of green vegetation (Dalposso et al., 2013). We would distinguish between different vegetation as well as the type of vegetation.
EBBI	$\frac{Band\ 5 - Band\ 4}{10\sqrt{Band\ 5 - Band\ 6}}$	The EBBI is a remote sensing index that applies wavelengths of 0.83 μm, 1.65 μm, and 11.45 μm, (NIR, SWIR, and TIR, respectively). These wavelengths were selected based on the contrast reflection range and absorption in built-up and bare land areas (As-syakur et al., 2012).
SAVI	$1.5 * \frac{Band\ 5 - Band\ 4}{Band\ 5 + Band\ 4 + 0.5}$	The SAVI minimizes the soil influence on vegetation quantification by introducing a soil adjustment factor (Polykretis et al., 2020).
SBI	$0.3037*Band2+0.2793*Band3+0.4743*Band4+0.5585*Band5+0.5082*Band9+0.1863*Band7$	Soil Brightness Index of TCT (Tasseled Cap Transformation) helps us to be able to distinguish bare soil (dry soil) and wet soil (Crist, 1985).
PCA1, PCA2, PCA3	PCA1 = a1,1Att1 +...+a1,NAttN PCA2 = a2,1Att1 +...+a2,NAttN PCAN = aN,1Att1 +...+aN,NAttN	PCA analysis was used with the main purpose is dimensionality reduction and exploring relationships between variables. In this study, 3 dimensions were selected with the most information

	are PCA1, PCA2, PCA3 (Brunsdon et al., 2012).
--	---

Tab. 2. Number of images in median composite, natural bands, indices bands and supporting bands

	No. images L8 in median composite	No. natural bands	No. indices bands	No. supporting bands (elevation)
Study site	14	6	9	1

Image segmentation and sample collection For satellite image classification, the image analysis was supported by the field trips in January 2020 and 2021 and the land use map in 2010 and 2020 to identify seven different types of Landcover. The first stage of the classification process is pixel-by-polygon-based image segmentation. The primary purpose of this process is to create uniformity in each image object. The multi-resolution segmentation algorithm in eCognition software is used in this study because it reduces the degree of disagreement of image objects for a specific resolution and is easy to implement based on the selection of the image objects. Select parameters about scale parameter, shape, and compactness. The scale parameter is the parameter that specifies the average size of the zoning result. This value can change due to the combination of many objects; it is indirectly related to the size of the created objects. The Scale value is proportional to the size of the object. The shape is a parameter with a variable value from 0-1, this value specifies the influence of colour on the segmentation process, and the Compactness parameter also has a value from 0-1, representing the texture in the object is created, the higher the value, the tighter the texture, the smaller the object. In this study, the values of the initial parameters (Scale parameter = 20, Shape = 0.5, Compactness = 0.7).

A total of 64524 segmented objects represent different overlay types: Barren land, Water, Built-up, Paddy, Orchard, Crops, and Forest. Adjacent object pairs are recombined to minimize heterogeneity. Objects with the same structure, shape, and tone are classified into the same category. However, in some regions, these objects fall into different categories. In contrast, there are also objects with different shapes and colour structures but included in the same category. Therefore, integrating visual interpretation with additional data collected in the field is essential to improve classification accuracy.

We carried out fieldworks in January, February 2020, and 2021 in Quang Tri and Thua Thien Hue to verify visual interpretation indoors: 40 polygons for each overlay type. Total 40 x 7 types = 320 polygons. The fieldwork in this study took place by comparing overlay data on 2016 Landsat 8 satellite imagery and 2020 Google Earth, and interviewing local managers to confirm that selected areas and Sampling have stability during 3-5 years. This helps the sampling data to be accurate, minimizing the effects in the process of land-use change. Figure 2 shows the overlay patterns on the Landsat 8 image. With the combination of natural colours, we can easily distinguish the water surface with its light tones and linear shapes. The density of vegetation is relatively high for natural forests, so these objects have a fairly smooth structure with natural light tones on satellite images. The bare soil has a light tone and a relatively uniform reflectance spectrum. The vegetation density is high for rice-growing, so the pixels have a relatively uniform reflectance spectrum with natural tones. Crops often have a bed-like structure mixed with ditches and have natural light tones. It is difficult to distinguish fruit trees and natural forests using pixel-based classification in the study area because they have relatively similar colour and shape structures. However, fruit trees usually have a much smaller area than natural forests. For residential areas, due to the appearance of many different objects such as houses, industrial zones, roads, bus stations, bridges, and gardens, they exhibit non-uniform spectral reflections and with relatively clear boundaries. The arrangement of objects in the residential area is regular, so they have similar tones and shapes. We showed details of the definition and description of land cover. Details on the feature set and the thresholds used, the rule set itself, and the hierarchy of objects are shown in Table 3.

Based on the analysis of the characteristics of the samples, the authors interpreted the types of land cover based on their colours, structures, and shape. The regions have the same colours, structures, and shapes combined with the same land cover types. In contrast, the regions have different colours, structures, and shapes, and the different types of land cover have been classified according to the authors' analyses. The regions have the same colours, structures, and shapes, but the natural characteristics are different; we used higher resolution images such as Google Earth and the land cover map to complete the analyses. The results of this step are significant for developing the following steps.

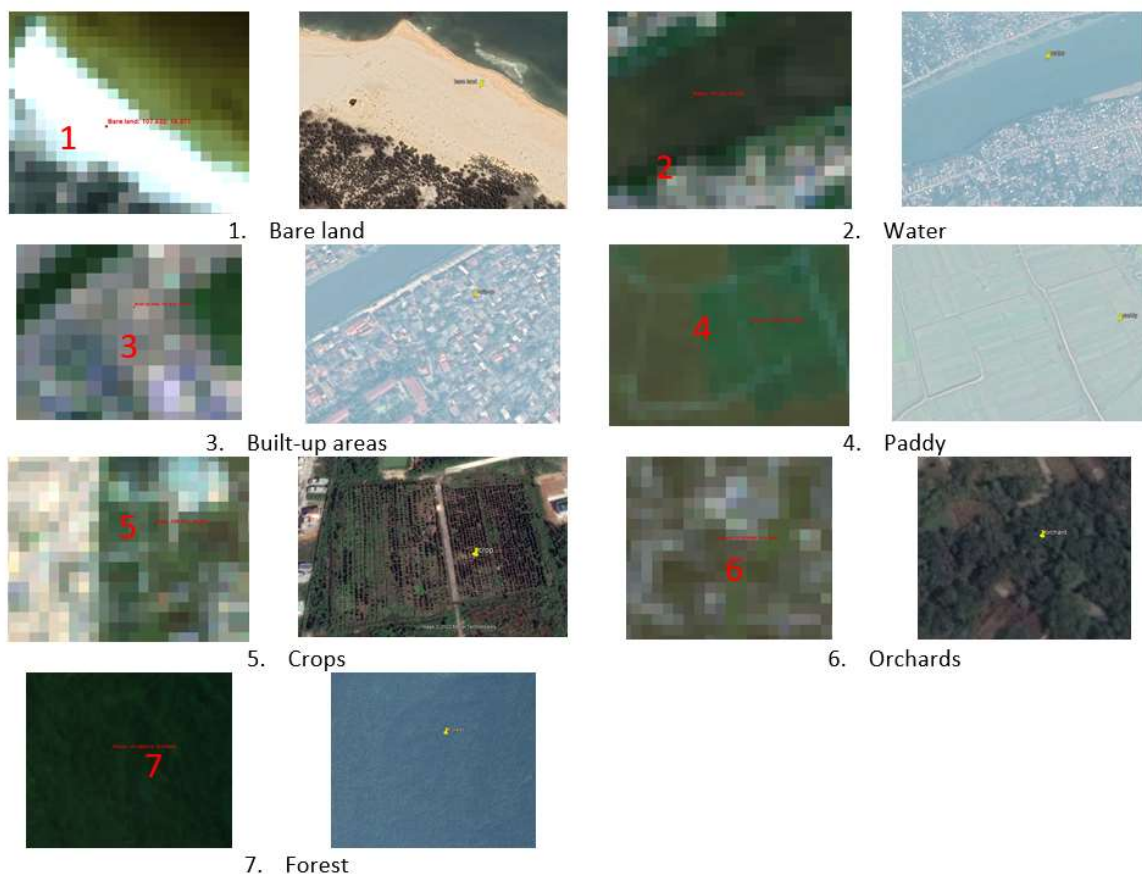



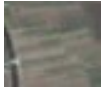



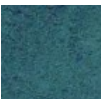


Fig. 2. Samples in the fields taken in the fields in 2/2020 and in the Google Earth in 2020: 1. Bare land (107.632055; 16.570518); 2. Water (107.560274; 16.455575); 3. Built-up area (107.573092; 16.47981); 4. Paddy (107.541403; 16.505351); 5. Crops (108.191181; 16.042115); 6. Orchards (107.576586; 16.434386); 7. Forest (107.554374; 16.218954).

Tab. 3. Definition land covers classes and description of the support region

Category	Definition	Description	Landsat Image	Google Earth image
Barren	Land with bare soil.	Barren include soils covered with sand, pebbles, rocks, stone surfaces, all other mineral materials or the bare ground in urban areas.		
Water	Land permanently covered with water or aquaculture sites with a water surface	These zones include submerged surfaces, always in the water. The limits are the banks or the vegetation without the coverage rate exceeding 25%.		
Built-up	Area covered with buildings or other types of construction (the object is characterized by a height and a texture composed of various natural or artificial materials). Land made impermeable by an artificial asphalt coating, concrete.	These areas include buildings of a permanent nature, covered with a roof (with some exceptions) intended to shelter, house or place people, animals, equipment and goods. These areas include all partially or fully waterproofed land, in particular asphalt, concrete, paved or slab floors. The paved road network, squares, car parks, etc., are waterproof non-building areas.		

Paddy	Paddy is the land planted the rice cultivation in the plain and the mountainous field.	Land is specialized in wet rice cultivation, other rice varieties and upland rice cultivation. The minimum area of the plots is 300m ² .		
Crops	Cropland is the annual cropland and annual cultivation land in the mountainous field.	Land for annual crops and vegetables or the plantation of flowers. The minimum area of the plots is 100m ² .		
Orchard	Orchard is perennial land on which there are groups of perennial trees for at least 3 years or more	Land for planting fruit trees such as palm, coconut, banana, litchi, longan, and industrial crops such as rubber, coffee, tea ...etc. Often trees will be planted in rows or around neighbourhoods. The minimum area of the plots is 500m ² .		
Forest	A forest is an ecosystem consisting of forest plants, animals, fungi, microorganisms, forest soil and other environmental elements, in which the main component is one or several species of trees and bamboo, cork, areca tree whose height is determined according to the flora on earthy mountains, rocky mountains, wetlands, sandy soil or other typical flora.	Area of inter-region from 0.3 ha or more; canopy level of 0.1 ha or more.		

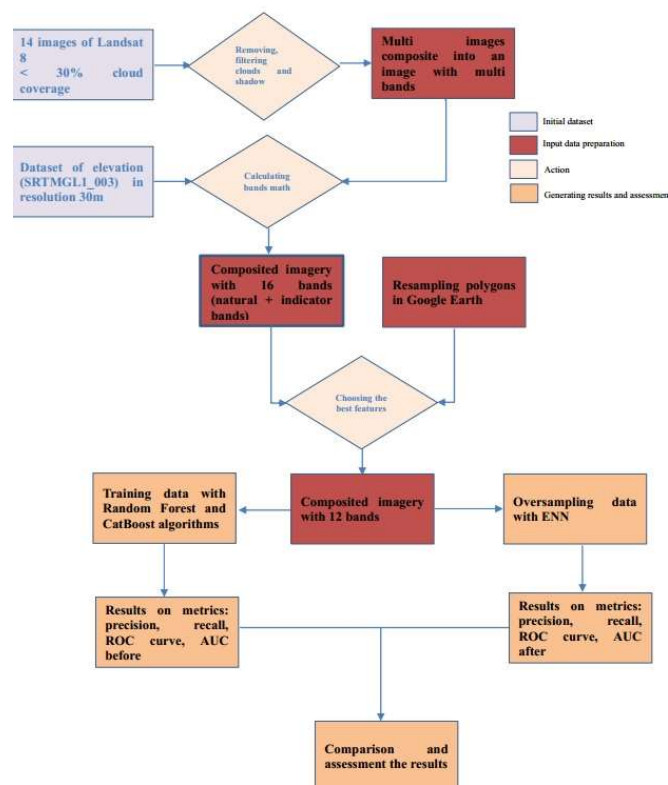


Fig. 3. Workflow of building the classification model

Methodology

The land cover classification methodology in this study was divided into six main steps (Figure 3): i) Removing and filtering the clouds with median composite images and using a quality assessment band; ii)

Extracting the information and detecting the objects in the image easily by calculating the indices bands; iii) Stacking all bands, normalizing and transforming all values. Numerical input variables may have a highly skewed or non-standard distribution. This could be caused by outliers in the data, multi-modal distributions, highly exponential distributions, and more. After that, applying feature selection to choose the best features or best bands with the most information for processing by mutual information statistics; Mutual information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. The scikit-learn machine learning library implements mutual information for feature selection with numeric input. This step is run on Google Colab; iv) Oversampling data with SMOTE ENN technique and training data with Random Forest and Catboost in Python; v) Assessing the model's performance; vi) Finally, assessing the results between existing datasets in Vietnam.

Random forest

Random Forest (RF) is a robust algorithm for solving classification and regression problems based on the results of decision tree prediction (Breiman 2001). RF allows combining many weak models to obtain the most influential models. Each sub-model (each decision tree) in the classification was evaluated using the majority voting method to determine the best methods (Dang et al., 2022). RF works in four main steps:

1. Resample from the bootstrap-based dataset to create subsets that have equal sizes to the original set.
2. Building a decision tree in a forest for each sample
3. Voting the prediction results
4. Returning the decision tree with the most votes

RF has the advantages of reducing bias and increasing model variance to avoid overfitting problems in the classification process (Mahdianpari et al., 2017). Also, the performance of the RF model depends on adjusting parameters like `max_features`, `n_estimators`, `min_sample_leaf`. As these parameters are directly related to the model's speed and prediction, the data set was separated in this study: 80% of the data were used as the bootstrap samples for each decision tree, while 20% of data was used to test the independent RF model. The number of trees and variables were tested in this study with 100, 500, and 1000. Finally, we chose Random Forest with 100 trees and tested it on two datasets: for using and non-using the SMOTE ENN technique.

Catboost

Cat Boost (CB) is a gradient boosting algorithm that applies the permutation technique to solve the classification problem. To reduce the overfitting problem, a random permutation and an average label value for samples with the same categories have been applied to the data set (Dorogush et al., 2018). CB works based on three main steps: i) the data set is randomly divided into subsets; ii) the label is converted to integers; iii) the encoding of the category features (Al Daoud 2019). In this study, CB parameters like Loss function, depth, and several iterations are refined to improve the performance of the classification model. The loss function has been set to "Multiclass." The depth was tested with the value of 5, 10, and 15; the best precision was 10. For the iterations parameter, after several refining times with the value of 300, 500, or 1000, the number of iterations is 500, which has the best accuracy using the statistical indices such as Receiver Operation Characteristics (ROC), the area under the ROC curve (AUC), Kappa (K), and Overall Accuracy (OA).

Feature selection

Sometimes the interplay between the bands and the classification model is complex. When we add more bands or indices to enhance the spectral value, this sometimes prevents the model's performance and accuracy. The input features would be a supervised procedure, and care must be taken to ensure that overfitting is not occurring. It is also desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in many cases, improve the model's performance. The feature selection aims to reduce the minor input features and avoid information redundancy. It yields the minimum generalization error (Vergara and Estévez, 2014). Mutual information (MI) measures statistical independence that has two main properties. First, it can measure any relationship between random variables, including nonlinear relationships (Thomas and Joy, 2006). Second, MI is invariant under transformations in the feature space that are invertible and differentiable, e.g., translations, rotations, and any transformation preserving the order of the original elements of the feature vectors (Kullback, 1997). We used the mutual information method to select features or bands before training data.

SMOTE ENN

Imbalanced data refers to datasets where several observations between two classes are very disparate. SMOTE-ENN is a hybrid sampling technique to fix the imbalance problem in the dataset, which Batista developed in 2004 (Batista et al., 2004). This method combines the SMOTE (Chawla et al., 2002) and ENN (Wilson, 1972). SMOTE is an over-sampling technique that can improve the model's classification accuracy by generating synthetic samples for the minority class, which means the minority class is over-sampled by increasing amounts.

The effect is to identify similar but more specific regions in the feature space as the decision region for the minority class. However, it could also create noise samples and boundary samples. The ENN technique will help overcome the disadvantage of SMOTE by removing observations from both classes that are identified as having different classes between the observation's class and its K-nearest neighbour majority class. SMOTE-ENN is applied in various fields such as Computer Science (Lu et al., 2019), Medical (Xu et al., 2020), Physical (Hassan et al., 2020), Financial (Aljawazneh et al., 2021), etc. and so on to solve problems that relate to the imbalanced data issue. The satellite imagery data has been composited by the median method in the time series and has been normalized and transformed into standard distributions of data. The process of SMOTE-ENN could be represented as (i) Initialization of SMOTE by selecting random observations from minority class; (ii) Determining the distance between the observations and its k nearest neighbours; (iii) Multiply the difference with a random number between 0 and 1 and put the result into the minority class as a synthetic sample; (iv) Repeat step (ii) and (iii) until the desired proportion of minority class is desiderated; (v) Initialization of ENN by defining K number; (vi) Identify the K-nearest neighbour of the observation among the other observations in the dataset and return the majority class from the K-nearest neighbour; (vii) Remove the observation and its K-nearest neighbour from the dataset if the class of the observation and the majority class from the observation's K-nearest neighbour is different; (viii) Repeat step (vi) and (viii) until the desired proportion of each class is fulfilled. All iterations are placed in a loop until the percentage between layers is balanced, then stops. In this study, Imbalanced-Learn implemented SMOTE-ENN, an open-source python library that relies on the Scikit-Learn package.

Accuracy Assessment

In this study, various statistical indices were used to evaluate the performance of LULC classification models, such as Receiver Operation Characteristics (ROC), the area under the ROC curve (AUC), Kappa (K), and Overall Accuracy (OA). ROC is a probability curve, and the area under the ROC curve reveals the reparability level of the model between classes. ROC curve represents the true positive rate (TPR) and false-positive rate (FPR). TPR is the sensitivity or proportion of the predictive class that was correctly detected, whereas FPR is the proportion that was incorrectly detected. In theory, the AUC value varies from 0 to 1, and 0.5 is considered a random threshold. The higher the TPR value, the higher the model's performance. At the same time, Kappa and the Overall accuracy index measure the differences between the prediction and observation values

Results

Exploration of the input dataset

Figure 4 shows the results of filtering and removing clouds. The clouds were filtered with the quality assessment band to obtain the transparent terrain pixel with the value of 2720.



Fig. 4. a. mean composite, b. median composite and c. median composite and clear terrain pixel filtering

The initial data we included are nine bands of natural spectrum, six indicators, and one supporting band is elevation. The composited bands are in the following order: Band2, Band3, Band4, Band5, Band6, Band7, EBBI, Elevation, EVI, GVI, NDWI, SAVI, SBI, PCA1, PCA2, PCA3. In Vietnam, there are regulations on specific land use plans for different types of land. Furthermore, the EVI, SBI, and GVI indices are enhanced indices for plant susceptibility or discriminating between luminances of different soils. Therefore, seasonal variation does not significantly affect the results. Furthermore, the study area has only two climates, the rainy season and the dry season. From May to September is between the two seasons, so the values obtained will be the median value. Sampling is combined with historical data on Google Earth, and when sampling needs to be collated and compared to get the sample median values. When we run the data to find the optimal attributes and give high accuracy, we find that the input attribute of 12 attributes is optimal (Figure 5).

>1 0.834 (0.006)
 >2 0.895 (0.008)
 >3 0.941 (0.007)
 >4 0.951 (0.007)
 >5 0.955 (0.006)
 >6 0.959 (0.005)
 >7 0.960 (0.006)
 >8 0.964 (0.005)
 >9 0.969 (0.004)
 >10 0.970 (0.005)
 >11 0.970 (0.005)
 >12 0.971 (0.005)
 >13 0.971 (0.005)
 >14 0.971 (0.005)
 >15 0.971 (0.005)
 >16 0.971 (0.004)

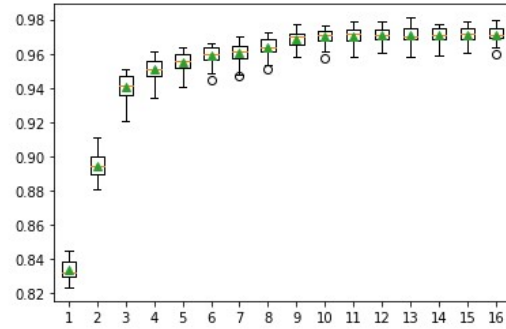


Fig. 5. Mean accuracy and standard deviation with the input features

The importance of the 12 variables selected in this study was evaluated using the Mutual Information method and incorporated with the Random Forest algorithm. The Feature 7 (elevation) is the most important variable with the important score value = 1.087736, which is followed by Feature 5: 1.065161 (B7), Feature 8: 1.045541 (EVI), Feature 6: 1.045404 (EBBI), Feature 11: 1.034742 (SAVI), Feature 4: 1.013988 (B6), Feature 9: 0.962317 (GVI), Feature 14: 0.922798 (PCA2), Feature 12: 0.882452 (SBI), Feature 13: 0.840174 (PCA1), Feature 2: 0.803828 (B4), Feature 3: 0.792662 (B5), Feature 0: 0.790812 (B2), Feature 10: 0.755551 (NDWI), Feature 15: 0.696559 (PCA3), Feature 1: 0.690080 (B3) (Figure 6).

Feature 0: 0.790812
 Feature 1: 0.690080
 Feature 2: 0.803828
 Feature 3: 0.792662
 Feature 4: 1.013988
 Feature 5: 1.065161
 Feature 6: 1.045404
 Feature 7: 1.087736
 Feature 8: 1.045541
 Feature 9: 0.962317
 Feature 10: 0.755551
 Feature 11: 1.034742
 Feature 12: 0.882452
 Feature 13: 0.840174
 Feature 14: 0.922798
 Feature 15: 0.696559

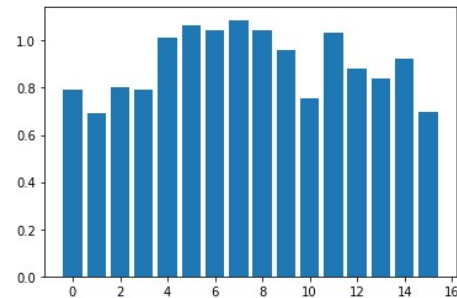


Fig. 6. The feature importance score

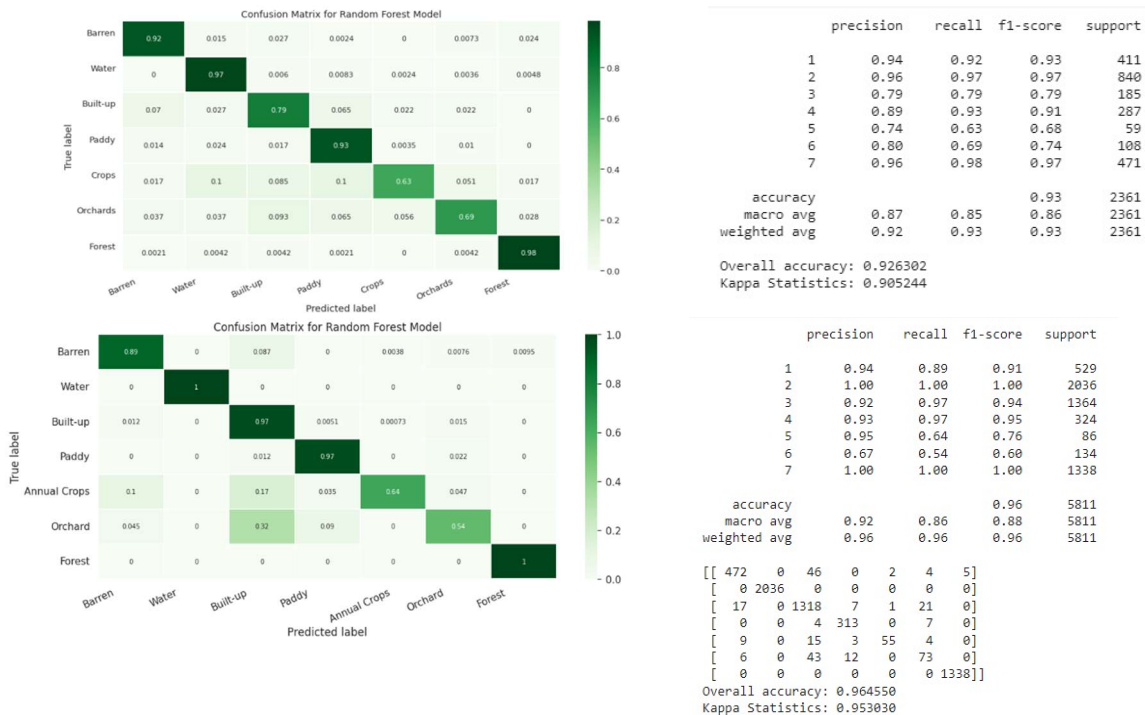


Fig. 7. Confusion matrix and Summary of statistics on the dataset before oversampling (1. Baren; 2. Water; 3. Built-up; 4. Paddy; 5. Crops; 6. Orchards; 7. Forest) – Random Forest (top); Catboost (down)

The accuracy assessment of classification results

Before oversampling, the data is imbalanced. We can see the bias between the classes. Class 5 and 6 (crops and orchards, respectively) have minor samples. Class 2 and 7 (water and forest, respectively) have the most examples and are biased towards the water and forest, which is indicated in the confusion matrix with the highest agreement in the diagonal of the matrix. In contrast, the crops and orchards class (class 5 and 6) has the lowest agreement (Figure 7).

Since the data has an imbalance, misclassification is inevitable. It is indicated in the summary of statistics above. Although the accuracy is high, the precision of the orchards, built-up, and crops class is low.

After applying SMOTE ENN technique, the samples in every class are balanced. The correct agreement in the matrix diagonal is almost the same in the confusion matrix. The precision is higher, which means the misclassification is reduced. We also find that the expected proportion of chance agreement after oversampling is higher than before oversampling (Kappa score before and after oversampling is 0.905244 and 0.974379 respectively for Random Forest and 0.953030 and 0.974379, respectively for CatBoost), which means the classifier after oversampling performed better than before oversampling when the data is balanced (Figure 8).

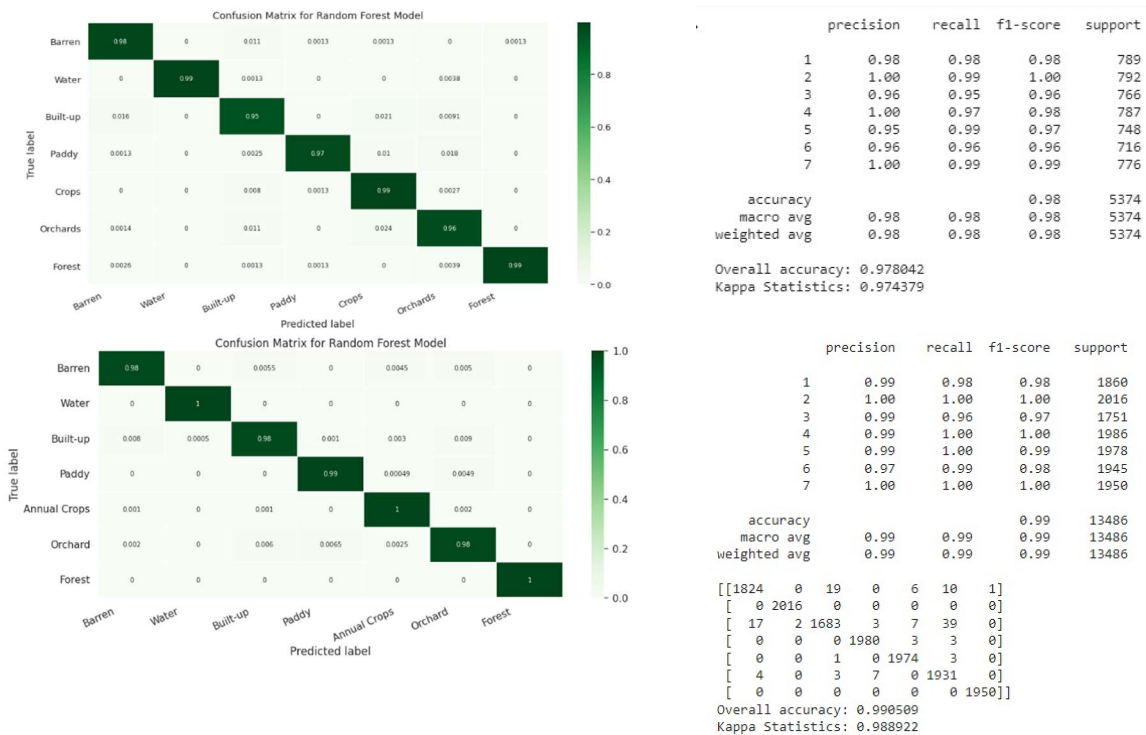


Fig. 8. Confusion matrices and Summary of statistics on dataset after oversampling (1. Baren; 2. Water; 3. Built-up; 4. Paddy; 5. Crops; 6. Orchards; 7. Forest) – Random Forest (top), CatBoost (dow)

Figure 9 shows the learning level of the model, and it found a clear difference between the model before and after oversampling. The precision-recall curve shows the relation between the accuracy and precision of the predictive model. The higher precision is, the lower the misclassification of the model is. Furthermore, the higher recall is, the lower the accuracy is. Both high scores show that the model returns accurate results. After oversampling, the precision increases, which means the misclassification of other instances in one class is decreased. Moreover, the recall is also higher, and the model's predictive capability is preciser.

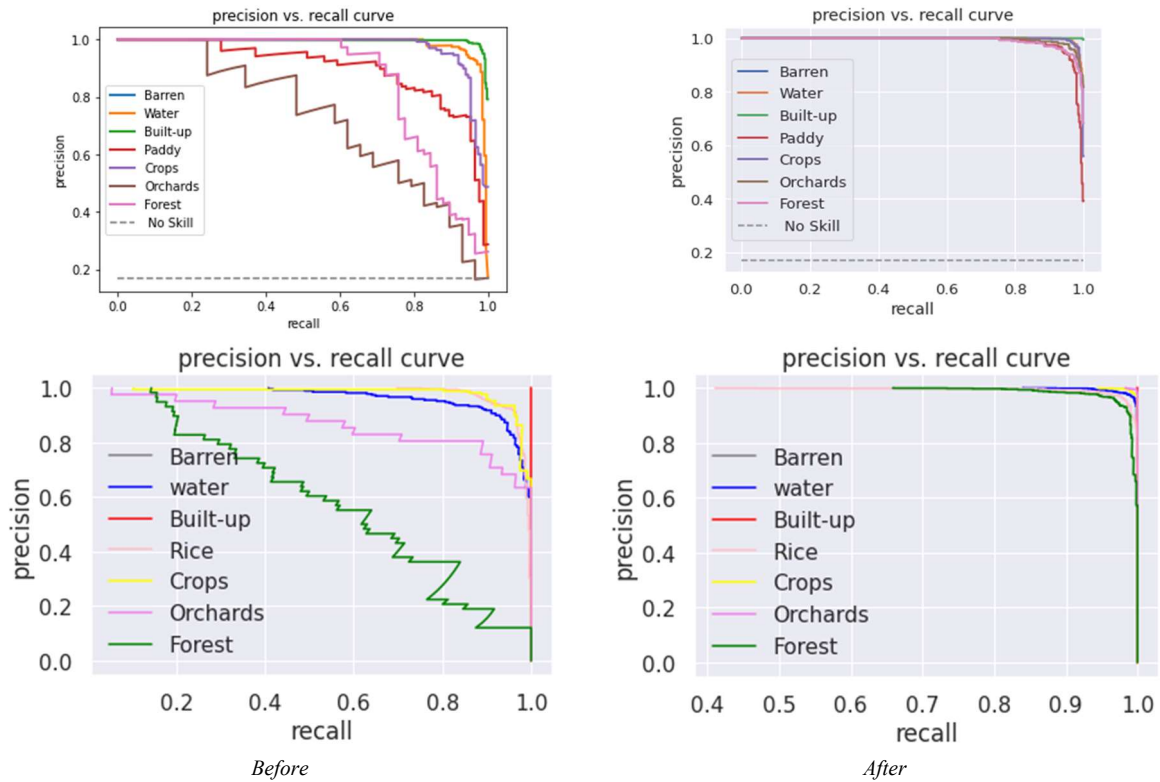
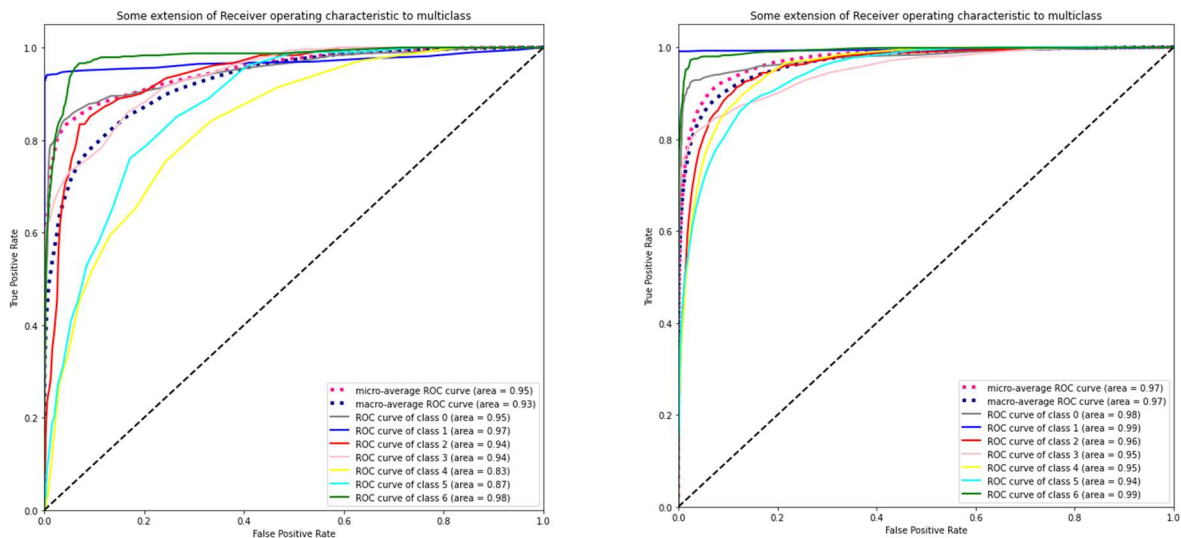


Fig. 9. The precision and recall curve for before oversampling and after oversampling – Random Forest (top), CatBoost (down)

To test the hypothesis about the strength of sample equilibrium, we can look at the classes' roc curve and AUC areas. Figure 10 shows that before oversampling, the prediction missing between types is significant, and after sample equalization, which decreases when the AUC areas between classes increase. We have tested two models of algorithms that give the same value.



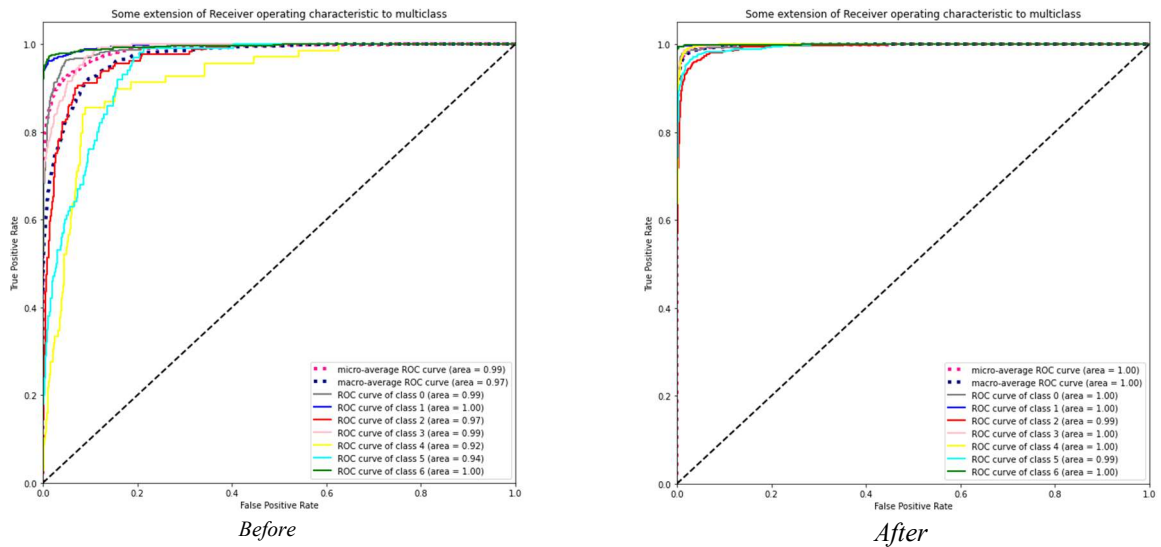


Fig. 10. Some extension of ROC - Random Forest (top) and CatBoost model (dow)
 Class 0: Barren; Class 1: Water; Class 2: Built-up; Class 3: Paddy; Class 4: Crops; Class 5: Orchards; Class 6: Forest

Based on the image classification results obtained, the samples are imbalanced (the samples of water are high while the examples of the crops and orchards are small, so the obtained classification image has very fragmented). Therefore, even paddy and crop areas are sometimes misclassified as water pixels. At the same time, classified images after equalizing samples have less fragmentation and confusion. Water pixels will be significantly reduced in addition to the fact that pixel crops and orchards are more accurately classified and generalized within the overall spatial distribution of the landcover.

Similarly, the CatBoost algorithm gave similar results when using the ENN method to balance the samples. The pixels of the landcover classes are also generalized better and are not fragmented in their total space (Figure 11).

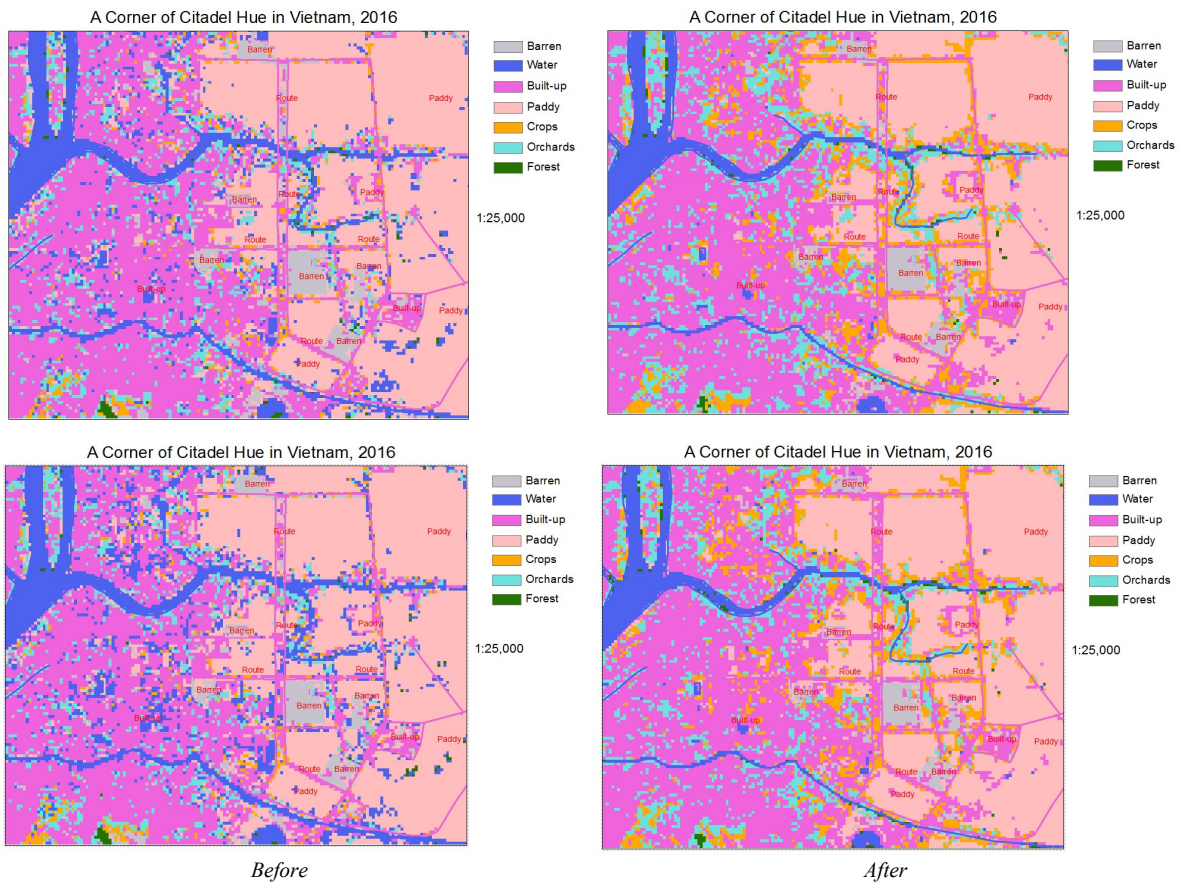


Fig. 11. Classification result of a corner Hue, 201 - Random Forest (top), CatBoost model (dow)

Discussion

Land managers and decision-makers need timely, accurate, and automated LULC information for sustainable land use planning (Burkhard et al., 2012, Pan et al., 2021). With the rapid development of remote sensing data and computing power, machine learning has been receiving the attention of global scientific communities in LULC classification (Bui et al., 2019a, Bui et al., 2019b). However, there are still several unresolved difficulties, and the data imbalance is considered one of these difficulties. Therefore, this study aims to develop a method to reduce the problem of imbalanced remote sensing data in near real-time at the national and regional scale in the classification of LULC.

In machine learning, the imbalance of the training data significantly affects the performance of the classification model, and several techniques have been proposed to solve this problem. Some have also been applied in experiments, such as generating more data by modifying the segmentation process's shape, compactness, and scale and generating data before the training stage (Bui et al., 2021). For example, the region of barren land covers this study area a smaller space in the study region; it takes up a small part of the whole training data. So, we changed the segmentation to generate smaller homogeneous objects to avoid data imbalance. However, this leads to a subdivision of all homogeneous objects in the satellite image and is inefficient. SMOTE ENN is considered one of the effective techniques to solve this problem. First, SMOTE calculates the difference between the characteristic vector (sample) under consideration and its closest neighbour, then multiplies this difference by a random number between 0 and 1 and adds it to the under consideration characteristic vector. Then the synthetic individuals are randomly scattered along the line between the minority class individual and its selected neighbours (Xu et al., 2020, Puri and Kumar Gupta, 2022). Thus, this approach makes the region of a decision of the minority class more prominent and more general.

Errors in LULC classification are inevitable; however, we can reduce them by applying more robust methods and techniques. Several studies have highlighted differing opinions on LULC classification accuracy. The classification accuracy depends on the quality of the satellite images, the characteristics of the topography, and the characteristics of the type of LULC. Anderson et al. (1976) reported that 85% accuracy for LULC classification is acceptable. (Jamali, 2020) had the accuracy of 88% when using Support Vector Machine and Complex Tree for LULC classification in Shiraz city, Iran, with Landsat 8 image. (Pan et al., 2021) applied Random Forest and CART for LULC classification in Australia and the USA with Landsat 5 imagery. Results showed 85-87% accuracy and 78-79% for Random forest and CART, respectively. (Qian et al., 2015) used support vector machine (SVM), normal Bayes (NB), classification and regression tree (CART), and K nearest neighbour (KNN) for LULC classification in the Haidian District of Beijing, China. The results showed that the accuracy varies from 70 to 95% depending on the land cover type. (Singh et al., 2021) machine learning (Mnlogit) to classify India's Landsat image in 2005, 2006, 2007, and 2011. The results highlighted that the accuracy ranges from 80-86% for all years. By applying the SMOTE ENN technique, the accuracy of the LULC classification in our study surpassed previous studies with an accuracy value of 98%.

There is always the demand to generalize the methods proposed for the different data and applications in machine learning. For example, the Random Forest and SMOTE ENN algorithm has proven effective in evaluating natural disasters such as floods, flash floods, or landslides. They have the effect of improving the LULC classification capability of this Vietnam case study. Specifically, as we mentioned above, there are some classification datasets in Vietnam (Some projects at the global scale, continental scale, or regional scale, such as Southeast Asia). Besides them, there is some researchers' dataset on the national scale. They have already worked on the MODIS images, Landsat incorporated with ALOS-2/PALSAR-2. Their classification productions are in resolutions 50m and 10m. Their method of resampling is different from our method. Although they went to in the field, they resample by points. Whereas we took by polygon, and our sample unit is the point. It means this method supports our algorithm to perform better. Their overall accuracies of classification production in 50m and 10m resolutions are 79.0% and 85.6%, respectively. Our overall accuracies are 92.6% and 97.8% (before and after oversampling).

The evidence is when we zoom into a corner of citadel Hue in Vietnam 2016. We can see the generalization of the classification and misclassification of the methods. With the dataset 50m, there is still misclassification between urban class and bare soil, even water class. We cannot detect the morphology of the river. With the classification production at 10m, the urban class (route) is misclassified into the crops class and bare soil. In addition, the river is misclassified into the orchards class. The classification image at a resolution of 10 m is misclassified the barren class into the built-up class and the misclassification between the crops and built-up (Figure 12).

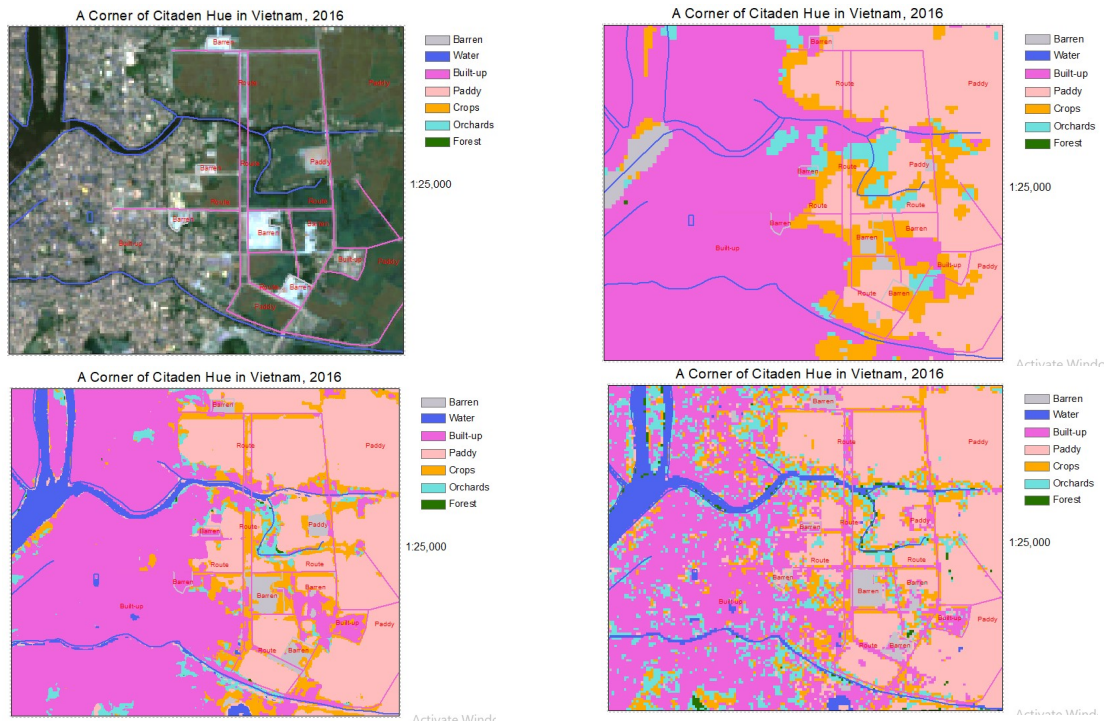


Fig. 12. a. Composition image in Hue, 2016; b. Classification result of a corner Hue in 2016 at 50m resolution; c. Classification result of a corner Hue in 2016 at 10m resolution; d. Classification result of a corner Hue in 2016 at 30m resolution (our production)

This study provides an essential and effective technique to monitor spatial and temporal changes in LULC in Vietnam and other countries worldwide. Our investigation can contribute to a framework for developing and elaborating strategies for sustainable land use planning. Although our study was applied for LULC classification, the finding can be used to evaluate environmental problems like floods, landslides, and flash floods

Conclusions

Monitoring LULC changes is a significant challenge for sustainable spatial planning, particularly in response to transformation and degenerative landscape processes. This study aims to develop a state-of-the-art technique based on machine learning, SMOTE ENN, and Remote Sensing for LULC classification in Vietnam.

The integration of oversampling and under-sampling with SMOTE ENN also improved the performance of the Random Forest and Casboost models for LULC classification. With Kappa's score before and after oversampling being 0.905244 and 0.974379, respectively, the new model can be used for classification in any region, especially in regions that are difficult to access and where data has been limited. Remarkably, this model can support decision-makers in developing strategies and appropriate policies for the planning and sustainable management of the territory.

Although this study successfully develops the LULC classification model in Vietnam, the construction of the models requires a faster process to better support the decision-maker in the monitoring process. Although this study has fully introduced the scientific implications and modelling processes, fully applying these processes is also a significant challenge in the local region. The results of this study can be an effective tool to analyze and monitor LULC change. However, these methods can be developed and assess other environmental problems

References

- Al Daoud, E. Comparison between xgboost, lightgbm and catboost using a home credit dataset. International Journal of Computer and Information Engineering, 2019, *Volume* 13 (1), pp. 6-10.
- Aljawazneh, H., Mora, A., García-Sánchez, P. & Castillo-Valdivieso, P. Comparing the performance of deep learning methods to predict companies' financial failure. IEEE Access, 2021, *Volume* 9, pp. 97010-97038.
- Aouissi, H.A., Petrișor, A.-I., Ababsa, M., Boștenaru-Dan, M., Tourki, M. & Bouslama, Z. Influence of land use on avian diversity in north african urban environments. MDPI: Land, 2021, *Volume* 10 (4), pp. 434.
- As-syakur, A. R., Adnyana, I. W. S., Arthana, I. W., Nuarsa, I. W. Enhanced built-UP and bareness index (EBBI) for mapping built-UP and bare land in an urban area. Remote Sensing, 2012, *Volume* 4(10), pp. 2957–2970.

- Batista, G.E., Prati, R.C. & Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 2004, *Volume 6* (1), pp. 20-29.
- Bhatti, S.S and Tripathi, N.K. Built-up area extraction using Landsat 8 OLI imagery. GIScience Remote Sens, 2014, *Volume 51*(4), pp. 445-467.
- Breiman, L. Random forests. Springer: Machine learning, 2001, *Volume 45* (1), pp. 5-32.
- Brennan, R. & Webster, T. Object-oriented land cover classification of lidar-derived surfaces. Canadian journal of remote sensing, 2006, *Volume 32* (2), pp. 162-172.
- Brunsdon, C., Harris, P., Fotheringham, A. S., McLoone, S., Demšar, U. Principal Component Analysis on Spatial Data: An Overview. Annals of the Association of American Geographers, 2012, *Volume 103*(1), pp. 106-128.
- Bui, Q.-T., Chou, T.-Y., Hoang, T.-V., Fang, Y.-M., Mu, C.-Y., Huang, P.-H., Pham, V.-D., Nguyen, Q.-H., Anh, D.T.N. & Pham, V.-M., 2021. Gradient boosting machine and object-based CNN for land cover classification. MDPI: Remote Sensing, 2021, *Volume 13* (14), pp. 2709.
- Bui, Q.-T., Nguyen, Q.-H., Pham, V.M., Pham, V.D., Tran, M.H., Tran, T.T., Nguyen, H.D., Nguyen, X.L. & Pham, H.M. A novel method for multispectral image classification by using social spider optimization algorithm integrated to fuzzy c-mean clustering. Canadian Journal of Remote Sensing, 2019a, *Volume 45* (1), pp. 42-53.
- Bui, Q.-T., Pham Van, M., Hang, N.T.T., Nguyen, Q.-H., Linh, N.X., Hai, P.M., Tuan, T.A. & Van Cu, P., 2019b. Hybrid model to optimize object-based land cover classification by meta-heuristic algorithm: An example for supporting urban management in ha noi, viet nam. International Journal of Digital Earth, 2019b, *Volume 12* (10), pp. 1118-1132.
- Burkhard, B., Kroll, F., Nedkov, S. & Müller, F. Mapping ecosystem service supply, demand and budgets. Ecological indicators, 2012, *Volume 21*, pp. 17-29.
- Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 200, *Volume 16*, pp. 321-357.
- Crist, E. P. SHORT COMMUNICATION A TM Tasseled Cap Equivalent Transformation for Reflectance Factor Data. Remote Sensing of Environment, 1985, *Volume 306*, pp. 301-306.
- Dang, K.B., Nguyen, T.H.T., Nguyen, H.D., Truong, Q.H., Vu, T.P., Pham, H.N., Duong, T.T., Nguyen, D.M., Bui, T.H. & Burkhard, B., 2022. U-shaped deep-learning models for island ecosystem type classification, a case study in con dao island of vietnam. One Ecosystem, 2022, *Volume 7*, pp. e79160.
- Dalposso, G. H., Uribe-Opazo, M. A., Mercante, E., Lamparelli, R. A. C. Spatial autocorrelation of NDVI and GVI indices derived from LANDSAT/TM images for soybean crops in the western of the State of Paraná in 2004/2005 crop season, Engenharia Agricola, 2013, *Volume 33*(3), pp. 525-537.
- Dorogush, A.V., Ershov, V. & Gulin, A. Catboost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363, 24 October 2018.
- Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. & Herrera, F. Learning from imbalanced data sets. Springer, 2018, pp. 79-121.
- Franks, S., Masek, J.G., Headley, R.M., Gasch, J. & Arvidson, T., Large area scene selection interface (lassi). Methodology of selecting landsat imagery for the global land survey 2005. Available online: <https://ntrs.nasa.gov/api/citations/20090027892/downloads/20090027892.pdf> (Access on 20 August 2020).
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, *Volume 42* (4), pp. 463-484.
- Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M. & Wolff, E. Very high resolution object-based land use-land cover urban classification using extreme gradient boosting. IEEE geoscience and remote sensing letters, 2018, *Volume 15* (4), pp. 607-611.
- Gutman, G., Byrnes, R.A., Masek, J., Covington, S., Justice, C., Franks, S. & Headley, R. Towards monitoring land-cover and land-use changes at a global scale: The global land survey 2005. Photogrammetric Engineering and Remote Sensing, 2008, *Volume 74* (1), pp. 6-10.
- Hassan, H., Ahmad, N.B. & Anuar, S., Year. Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mininged. Journal of Physics: Conference Series IOP Publishing, 052041.
- Huang, C., Davis, L. & Townshend, J. An assessment of support vector machines for land cover classification. International Journal of remote sensing, 2002, *Volume 23* (4), pp. 725-749.
- Jamali, A. Land use land cover mapping using advanced machine learning classifiers: A case study of shiraz city, iran. Earth Science Informatics, 2020, *Volume 13* (4), pp. 1015-1030.
- Kadavi, P.R. & Lee, C.-W., 2018. Land cover classification analysis of volcanic island in aleutian arc using an artificial neural network (ann) and a support vector machine (svm) from landsat imagery. Geosciences Journal, 2018, *Volume 22* (4), pp. 653-665.
- Kullback, S. Information theory and statistics. Dover Publications, New York, 1997, p. 409.

- Lambin, E.F. & Geist, H.J. Land-use and land-cover change: Local processes and global impacts: Springer Science & Business Media, 2006, p. 236
- Lambin, E.F., Rounsevell, M.D. & Geist, H.J. Are agricultural land-use models able to predict changes in land-use intensity? Agriculture, Ecosystems & Environment; ELSEVIER, 2000, *Volume* 82 (1-3), pp. 321-331.
- López, V., Fernández, A., García, S., Palade, V. & Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 2013, *Volume* 250, pp. 113-141.
- Lu, T., Huang, Y., Zhao, W. & Zhang, J., Year. The metering automation system-based intrusion detection using random forest classifier with smote+ enn. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)IEEE, 19-20 October 2019, pp. 370-374.
- Mahdianpari, M., Salehi, B., Mohammadimanesh, F. & Motagh, M., 2017. Random forest wetland classification using alos-2 l-band, radarsat-2 c-band, and terrasars-x imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, *Volume* 130, pp. 13-31.
- Mohamed, M.A., Anders, J. & Schneider, C. Monitoring of changes in land use/land cover in Syria from 2010 to 2018 using multitemporal Landsat imagery and GIS; MDPI: Land, 2020, *Volume* 9 (7), p. 226.
- Pan, X., Wang, Z., Gao, Y., Dang, X. & Han, Y. Detailed and automated classification of land use/land cover using machine learning algorithms in google earth engine. Geocarto International, 2021, pp. 1-18.
- Petrişor, A.-I., Sirodoev, I. & Ianoş, I. Trends in the national and regional transitional dynamics of land cover and use changes in romania. MDPI: Remote Sensing, 2020, *Volume* 12 (2), p. 230.
- Platt, R.V. & Rapoza, L. An evaluation of an object-oriented paradigm for land use/land cover classification. The Professional Geographer, 2008, *Volume* 60 (1), pp. 87-100.
- Polykretis, C., Grillakis, M. G., Alexakis, D. D. Exploring the impact of various spectral indices on land cover change detection using change vector analysis: A case study of Crete Island, Greece. Remote Sensing, 2020, *Volume* 12(2).
- Prakasam, C. Land use and land cover change detection through remote sensing approach: A case study of kodaikanal taluk, tamil nadu. International journal of Geomatics and Geosciences. International Journal of Geomatics and Geosciences, 2010, *Volume* 1 (2), p. 150.
- Puri, A. & Kumar Gupta, M. Improved hybrid bag-boost ensemble with k-means-smote-enn technique for handling noisy class imbalanced data. The Computer Journal, 2022, *Volume* 65 (1), pp. 124-138.
- Qian, Y., Zhou, W., Yan, J., Li, W. & Han, L. Comparing machine learning classifiers for object-based land cover classification using very high-resolution imagery. Remote Sensing, 2015, *Volume* 7 (1), pp. 153-168.
- Reidsma, P., König, H., Feng, S., Bezlepkina, I., Nesheim, I., Bonin, M., Sghaier, M., Purushothaman, S., Sieber, S. & Van Ittersum, M.K. Methods and tools for integrated assessment of land use policies on sustainable development in developing countries. Land Use Policy, 2011, *Volume* 28 (3), pp. 604-617.
- Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS journal of photogrammetry and remote sensing, 2012, *Volume* 67, pp. 93-104.
- Shao, Y. & Lunetta, R.S. Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points. ISPRS Journal of Photogrammetry and Remote Sensing, 2012, *Volume* 70, pp. 78-87.
- Singh, R.K., Singh, P., Drews, M., Kumar, P., Singh, H., Gupta, A.K., Govil, H., Kaur, A. & Kumar, M. A machine learning-based classification of landsat images to map land use and land cover of india. Remote Sensing Applications: Society and Environment, 2021, *Volume* 24, 100624 Available from: <https://www.sciencedirect.com/science/article/pii/S2352938521001609> (Access on 15 December 2021).
- Song, X., Duan, Z. & Jiang, X. Comparison of artificial neural networks and support vector machine classifiers for land cover classification in northern china using a spot-5 hrg image. International Journal of Remote Sensing, 2012, *Volume* 33 (10), pp. 3301-3320.
- Thanh Noi, P. & Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. Sensors, 2018, *Volume* 18 (1), p. 18.
- Thomas, M. & Joy, A.T. Elements of information theory. Wiley-Interscience, 2006, p. 774
- Turner, W., Rondinini, C., Pettorelli, N., Mora, B., Leidner, A.K., Szantoi, Z., Buchanan, G., Dech, S., Dwyer, J. & Herold, M. Free and open-access satellite data are key to biodiversity conservation. Biological Conservation, 015, *Volume* 182, pp. 173-176.
- Vergara, J.R. & Estévez, P.A. A review of feature selection methods based on mutual information. Neural computing and applications, 2014, *Volume* 24 (1), pp. 175-186.
- Vluymans, S. Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods. Springer, 2019, *Volume* 107, p. 236.
- Wilson, D.L. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, 1972, *Volume* (3), pp. 408-421.

- Xu, Z., Shen, D., Nie, T. & Kou, Y., 2020. A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data. *Journal of Biomedical Informatics*, 2020, *Volume* 107, p. 103465.
- Yijing, L., Haixiang, G., Xiao, L., Yanan, L. & Jinling, L. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 2016, *Volume* 94, pp. 88-104.