BỘ NÔNG NGHIỆP VÀ PTNT
TẠP CHÍ NÔNG NGHIỆP VÀ PTNT
***

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
-----------***-----------

Hà Nội, ngày 29 tháng 6 năm 2023

GIẤY XÁC NHẬN

Tạp chí Nông nghiệp và Phát triển Nông thôn là cơ quan ngôn luận của Bộ Nông nghiệp và PTNT; tuyên truyền các chủ trương của Đảng, chính sách pháp luật của Nhà nước về lĩnh vực nông nghiệp và phát triển nông thôn; là diễn đàn khoa học và công nghệ, trao đổi kinh nghiệm thực tiễn, lý luận, nghiệp vụ về nông nghiệp và phát triển nông thôn.

Toà soạn Tạp chí Nông nghiệp và Phát triển Nông thôn xác nhận đã nhận được bài báo: "Applying random forests algorithm for land cover mapping based satellite Imagery" *của tác giả* Tran Thi Hoa [1,*]

[1] Faculty of Information Technology, Hanoi University of Mining and Geology

Bài báo đã được duyệt đăng trên Tạp chí Nông nghiệp và Phát triển nông thôn bằng ngôn ngữ Tiếng Anh.

K.T TỔNG BIÊN TẬP
PHÓ TỔNG BIÊN TẬP

BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG THÔN
TẠP CHÍ
NÔNG NGHIỆP VÀ
PHÁT TRIỂN
NÔNG THÔN
★

Dương Thanh Hải

Applying Random Forests Algorithm for Land Cover Mapping based Satellite Imagery

Hoa Thi Tran

Hanoi University of Mining and Geology

ABSTRACT:

Landcover mapping is vital for comprehending Earth's surface features, encompassing both natural (vegetation, hydrological systems) and man-made elements. It provides essential information for sustainable and responsible land management practices, helping to balance economic development with environmental conservation. There are several approaches to mapping land cover, such as: field surveys, geospatial ananlysis, or remote sensing. Each of these methods has its own strengths and limitation depending on factors such as the scale of mapping, availability of data, accuracy requirements, and budgetary constraints. In order to achieve the most accurate and detailed results, landcover mapping often involves a combination of these methods. This study proposed one of common machine learning algorithms – random forests to be used to classify land cover types automatically. This algorithm learned patterns and features from training data collected from a Landsat 8 scence of Phu Ly city, Ha Nam province, then applied them to classify unlabeled data of the whole scence. The approach yielded a remarkable 95% accuracy, surpassing alternatives such as a popular maximum likelihood supervised classfication. Accurate land cover mapping facilitates decision-making, assesses land use changes, and supports sustainable land management. It provides valuable insights for environmental monitoring, urban planning, and biodiversity conservation. Thus, the random forests approach has shown promisisng results in land cover mapping, enhancing our understanding of Earth's dynamic landscape.

Keywords: *land cover mapping, Landsat8, machine learning, maximumlikelihood, random forests.*

1. Introduction

Land is one of the most important natural resources; is a home for all living species and human; and is place to host all physical, climatic and economic activities [1]. Land cover is referring to the surface covering over on the ground such as shrubland, built-up, water or vegetation, etc. The other term sharing similarities to land cover is land use, which refers to purposes of the assigned land, like agriculture, forestation, range land. Land use and land cover (LULC) commonly cohere to illustrate both humane activities and natural elements on the landscape. Industrialization, economic growth and urbanization are processes that mainly result changes the landscape over a specific time frame. Thus, it is important to conduct a better understanding of how land has been utilized as well as an involvement of land management policies and land monitoring to ensure sustainable development [2]. Mapping LULC over a specific time frame carries out a responsibility of supporting materials for land management by facilitating resource allocation, environmental monitoring, land use planning, natural resource management, and disaster risk assessment planning.

There are several methods available for landcover mapping. Remote sensing involves using satellite or aerial imagery along with image classification algorithms to identify and classify different land cover types [3] [4]. Field surveys involve direct observations and data collection on the ground to validate and improve landcover maps [5] [6]. Geospatial analysis combines various spatial datasets, such as satellite imagery and GIS data, to generate or update landcover maps [7]. LiDAR technology utilizes laser beams to measure surface features, such as topography and vegetation structure [8] [9]. Data fusion combines information from multiple sensors or sources to enhance the accuracy and detail of landcover maps [10] [11]. Machine learning and artificial intelligence techniques use algorithms to automatically classify land cover based on training data [12] [13]. Unsupervised and supervised classification methods group pixels or objects in an image based on their spectral characteristics. The choice of method depends on factors such as scale, data availability, accuracy requirements, and budget constraints. Often, a combination of these methods is employed to achieve accurate and comprehensive landcover mapping results [14] [15].

In this study, we investigated in a collarboration of using satelite image and one of machine learning methods – the random forests. Remotely sensed data are common resources for LULC mapping programs because of their advantages of providing many prospects to obtain physical statuses of LUCL at a certain or various spatial and temporal resolution [16]. Landsat images are more appropriate for mapping LULC at moderate scale (level 1 or level 2 of land classification systems respectively) within up to 90% of accuracy expectation of the Maximum Likelihood method (ML)– a very widely used technique of the image classification process [16]. However, there are other prospective methods that allows to get a higher accuracy of the analysis at some specific cases, such as "Random Forests technique" (RF) which uses a hierarchy of decision tree to assign samples into each class [17]. Therefore, we also examined how effective the random forests and ML supervised classificarion were in our case study. However, it is important to underline these factors in order to extract land cover information from satellite images: (1) a level of land classification system; (2) a requirement of accuracy; (3) a type of a chosen image; and (4) an image classification method.

Our study area is the entire Phu Ly city of Ha Nam Province where rapid urbanization has been occurring. The pace of urban expansion in the area has led to an increased demand for frequent updates on land status and comprehensive insights into land management for local authorities. As the city experiences significant growth and development, it becomes crucial to have up-to-date information on land cover and land use patterns. Accurate and timely landcover mapping is essential to understand the dynamics of urbanization, monitor changes in land use, and support effective land management decisions. By providing a comprehensive understanding of the current land status, including the distribution of urban, agricultural, residential, and other land cover types, our study aims to assist local authorities in making informed decisions, planning for infrastructure development, and ensuring sustainable land management practices in Phu Ly,

2. Methodologies

2.1 The process overview

The methodologies employed in this study utilizes Landsat 8 imagery and the random forest classification algorithm for land cover mapping. The following steps outline the approach (figure 1 highlighted the whole process in a flowchart):

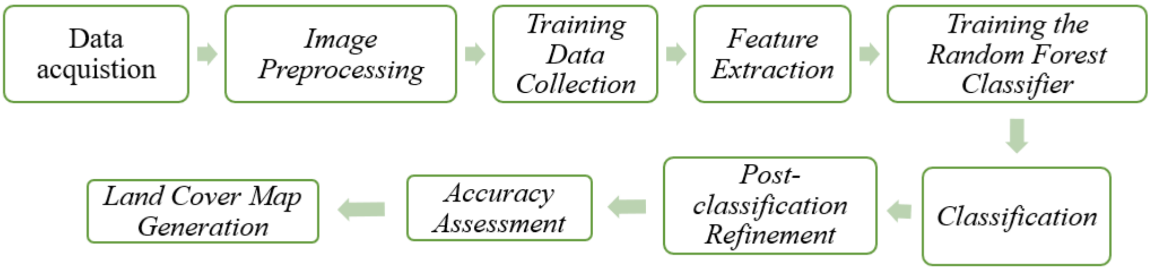


Figure 1. The flowchart decribled the 9 steps in mapping land cover using Landsat 8 scence and random forest of the study area.

Step 1- Data Acquisition: Landsat 8 imagery is acquired for the study area, Phu Ly city of Ha Nam Province, to capture the necessary spectral information required for land cover mapping. In this study, we acquired one Landat scence of Phu Ly with the ID number was LC08_L2SP_127046_20210717_20210729 from webiste: https://earthexplorer.usgs.gov/.

Step 2- Preprocessing: The Landsat 8 imagery is preprocessed to correct for atmospheric effects, sensor artifacts, and radiometric calibration. This ensures the data is in a suitable form for analysis.

Step 3- Training Data Collection: Ground-truth data or reference points representing different land cover classes are collected through field surveys or existing land cover datasets. These reference points serve as training data for the random forest classifier.

Step 4- Feature Extraction: Relevant spectral, textural, or spatial features are extracted from the preprocessed Landsat 8 imagery to characterize the different land cover classes. These features capture the distinguishing characteristics of each land cover type.

From step 2 to step 4, we worked on Erdas software then summarized the data as checkpoints for the next steps of training and classifying in CART Navigator - Salford.

Step 5- Training the Random Forest Classifier: The extracted features from the training data are used to train the random forest classifier. The algorithm learns the relationships between the spectral signatures and the corresponding land cover classes. Further information of this method was represented in the following section.

Step 6- Classification: The trained random forest classifier is applied to the entire Landsat 8 image, classifying each pixel into one of the predefined land cover classes based on its spectral characteristics. The classifier assigns a probability or confidence level to each class to quantify the uncertainty of the classification. In section 2.3, we dicussed further the system applied in land cover classification.

Step 7- Post-classification Refinement: Post-classification techniques, such as spatial filtering or object-based analysis, can be employed to refine the land cover map, improve accuracy, and reduce classification errors.

Step 8- Accuracy Assessment: The accuracy of the land cover map is assessed by comparing the classified results with independent reference data. This evaluation helps quantify the reliability and overall accuracy of the classification. We also conducted a set of training samples for the Maximum Likelihood to compare to our selected random forests.

Step 9- Land Cover Map Generation: The final output is a land cover map that provides information about the distribution and spatial extent of different land cover classes within the study area, Phu Ly city.

2.2.Random Forests method

Random Forests (RF) is a machine learning algorithm to reduce number of training data relatively based on defined parameters [17]. The regression of training samples' numbers is practically depending on how to split the decision tree nodes. Decision trees generally are models constructed by a set of binary rules to estimate (predict and calculate) a target value [16]. Figure 4 described how samples were assigned in decision trees and classified into class. There are two types of decision trees: regression tree (figure 2) and classification that are normally as known as CART (Classification and Regression Tree). However, in RF method, there will be "n" CART that supports the whole process within less supervision of analysers [16].

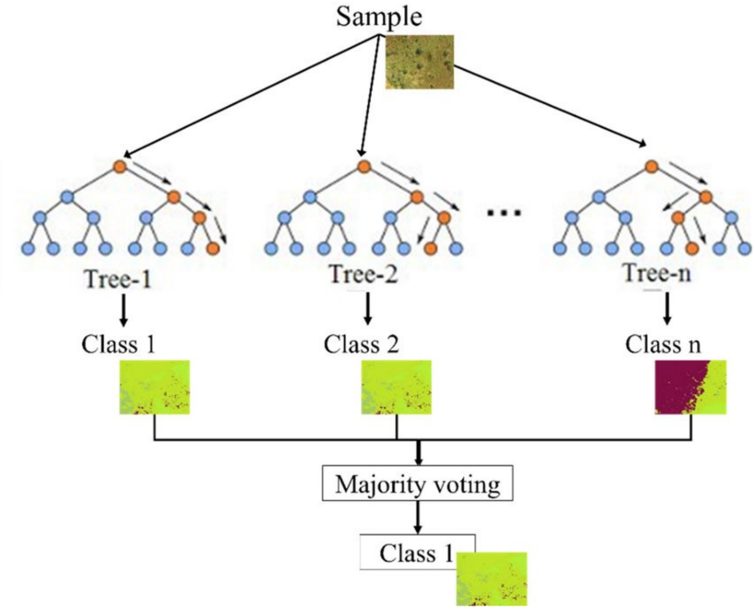


Figure 2. A general hierarchy of decision trees of the RF for classification [18]

2.3.Land cover classification system

As followed by FAO ([1], the land cover classification system (LCCS) involves two phases: "Dichtomous" and "Modular-Hierarchical" phase, where:

- The Dichitomous phase defines eight types of landcover: (1) Cultivated and Managed Terrestrial Areas; (2) Natural and Semi-Natural Terrestrial Vegetation; (3) Cultivated Aquatic or Regularly Flooded Areas; (4) Natural and Semi-Natural Aquatic or Regularly Flooded Vegetation; (5) Artificial Surfaces and Associated Areas; (6) Bare Areas; (7) Artificial Waterbodies, Snow and Ice; and (8) Natural Waterbodies, Snow and Ice.

- The Modular – Hierarchial phase designates land cover classes deriving from those 8 major pre-defined land cover types above, resulting a system of land cover types including:

- a *Boolean formula* showing each classifier used (all classifiers are coded);
- a *unique number* for use in Geographical Information Systems (GIS); and
- a *name*, which can be the standard name as supplied or a user-defined name.

However, in this experiment, the chosen system is simplified from the FAO instruction based on the ideal of the map scale and the geographic area to determine the land types (legend) and mapping units (see figure 3). Thus, we chose the first level of land cover type classification, which are: vegetation, urban, open water, and others.

- *Vegetation or vegetated area*: where plants and trees mainly dominate;

- *Urban or residential areas*: building and industrial infrastructure;

- *Open water*: natural and artificial hydrological systems (lakes, ponds, rivers, streams);

- *Others or (sometime unclassified):* roads, streets, bare soils.

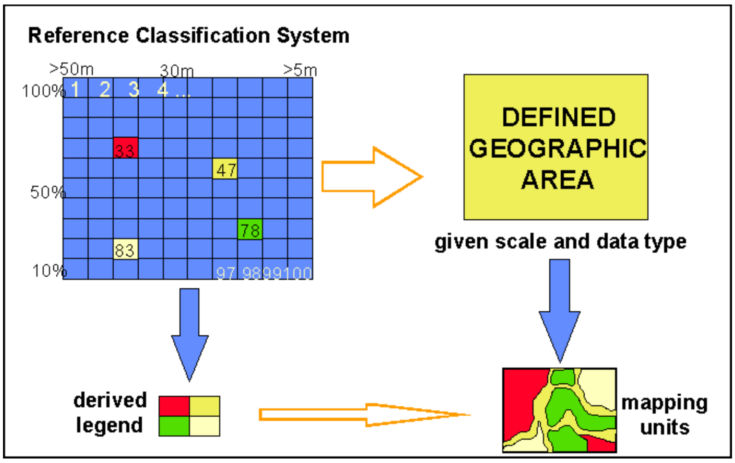


Figure 3. The FAO classification systems [1]

3. Results and Discussion

3.1. Data input

Data using to conduct experimental results are a Landsat 8 OLI image captured Phu Ly city of Ha Nam province in July 2021, local maps and other records for sampling process and accuracy assessment.

The Landsat 8 image was pre-processed to correct atmospheric and geometric effects, and to convert radiance value of each pixel to digital number (DN from 0-255) for the next step of the processing. Figure 4 shows the “false-color” composite image (5-4-1) in which different land cover types appears in various color scheme: bright green for vegetation, dark blue for open water and brownish tan for urban. Figure 5 is an extracted NDVI (Normalized Difference Vegetation Index) image after the pre-processing step, representing pixel values from -1 to 1. In LULC mapping, NDVI index is mainly used as one of indicators for selecting samples.

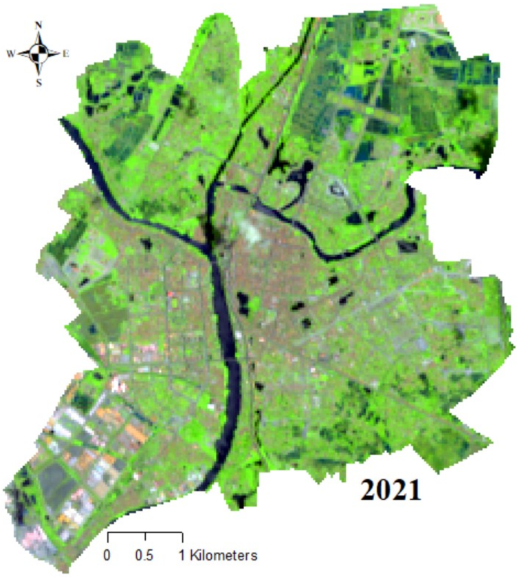


Figure 4. A Landsat composite image: vegetation (green), open water (dark blue); urban (brown).

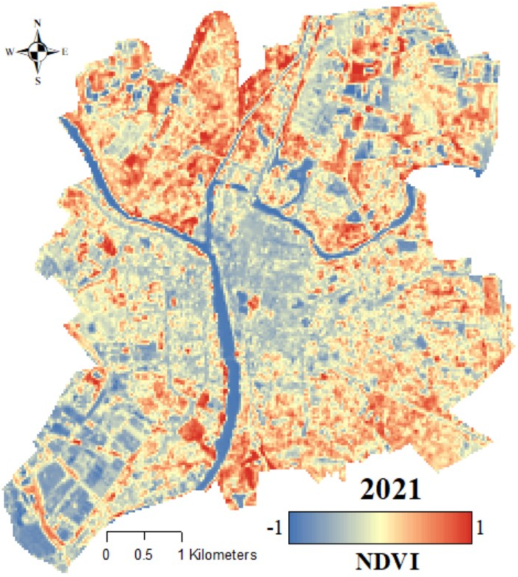


Figure 5. A NDVI image generated after processing, values are from -1 to 1

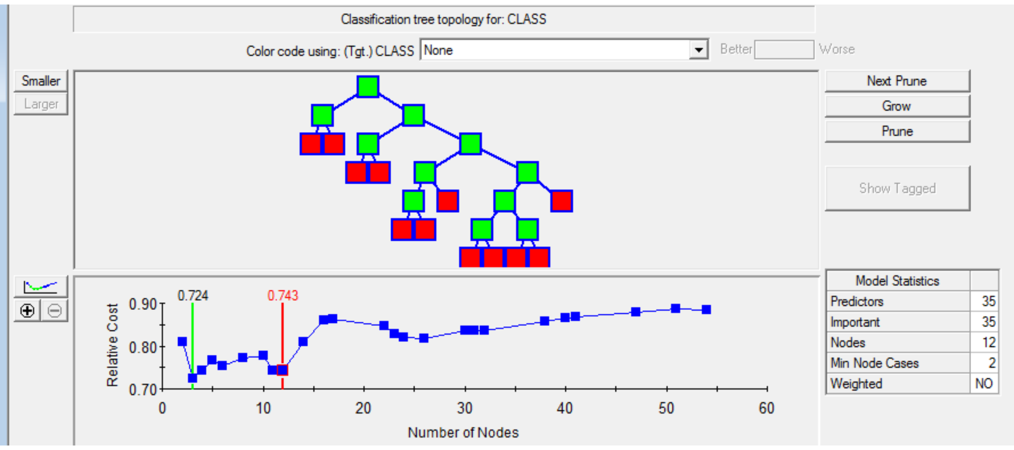
3.2. Traning data and class extraction

We created totally 413 data points. Firstly, the surface reflectance of all bands for each checked point was extracted and converted to an ASCII file format. This file was then converted to an Excel file (.xls) format for compatibility with the Salford software. The Excel file contained approximately 413 points with indexes of band reflectance, class, and XY-coordinates.

Next, the class variable was chosen as the target variable, and the band reflectance values served as predictors for testing the prediction success. A cross-validation process was conducted using the 413 values to assess the accuracy of the predictions.

The random forest algorithm was applied using the Salford software, which automatically generates decision trees with approximately 54 nodes. These decision trees had different relative costs and collectively contribute to the land cover mapping process.

By utilizing these steps, the land cover mapping process leveraged the band reflectance values as predictors and the class variable as the target, allowing for the generation of decision trees that aid in accurately classifying the land cover types within the study area. Firgue 5 represnted the CART at node 12 and 0.743 relative cost respectively.



Classification tree topology for: CLASS

Color code using: (Tgt.) CLASS None

Model Statistics	
Predictors	35
Important	35
Nodes	12
Min Node Cases	2
Weighted	NO

Figure 5. CART at node 12

Figure 6 showes a detailed example of how a classification tree applied in the experiment using red ban (band 3) and near infrared band (band 4) as parameters. The top of the tree is the "root node" evaluating the rule of "values in band 4 less than or equal to 40" will be designated into "water". Other decisions will need other rules until all branches end which means all pixels are assigned into land cover types.

In this experiment, we used seven bands of a Landsat 8 image as nodes to construct decision trees due to the given conditions of moderate spatial resolution and level 1 of land classification system (4 land cover types). However, the nodes and trees will grow if there are more land cover types involved at more complicated classification system.

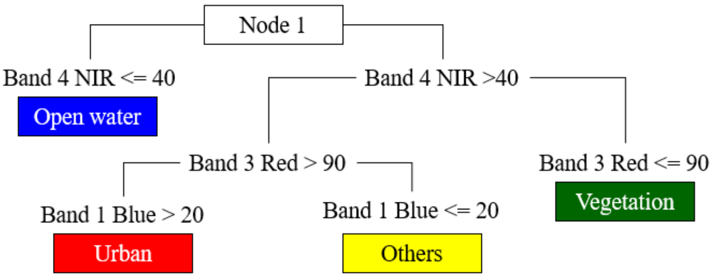
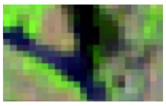
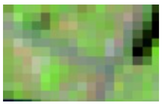


Figure 6. An example of Node 1 in a decision tree to assign the samples into a class

We also conducted a Maximum Likelihood classification (ML) comparing how effective the RF method wass. ML method is commonly used in image classification assigning pixels with a high potential similarity into class. The same training samples collection was carried out for both algorithms showed in table 1.

Table 1. Training samples for two classifications methods, RF and ML.

Land cover type	Indicators	Samples
Urban	- Shape: rectangles, blocks; - Color: brown, tan or gray; - Distribution: convergent or discreted.	

Vegetation	- Shape: non-homogenous; - Color: light green to dark green; - Distribution: convergent.	
Open water	- Shape: linear or non-homogenous, somes are small size from 10-20 pixels; - Color: blue, dark blue to black; - Distribution: locating along side the boundary, interspred in urban areas.	
Others (a mix of roads and bareland)	- Shape: linear, non-homogenous; - Color: greenish gray; - Distribution: discreted or convergent around intersection;	

3.3. Discussion

Our results showed that in both methods, there was no un-classified land type. All pixels were assigned into desiring classes. Maps of land cover are illustrated in figure 7 to the RF method and figure 8 to the ML one. Generally, urban areas were covering more than 40% of the total while open water was less than 10%. In the RF method, roads and streets were captured more accurate of which occurence was more obviously linear (class of "other"). In the ML method, on the other hand, those road features scattered around urban areas without homogenous shapes resulting the indicator "shape" did not work well. It is understandable because the ML statistical algorithm is mainly based on spectal information of pixel. Therfore, separativity wasnot an easy step depending on the quality and condition of capturing images. This experienced image was collected in July 2021 which was in the rainy season, so soil was more moisturous leading to some lands could be assigned into "open water", for example of the lands on the south west of reseacher area on ML classified map while practically they were urbans.

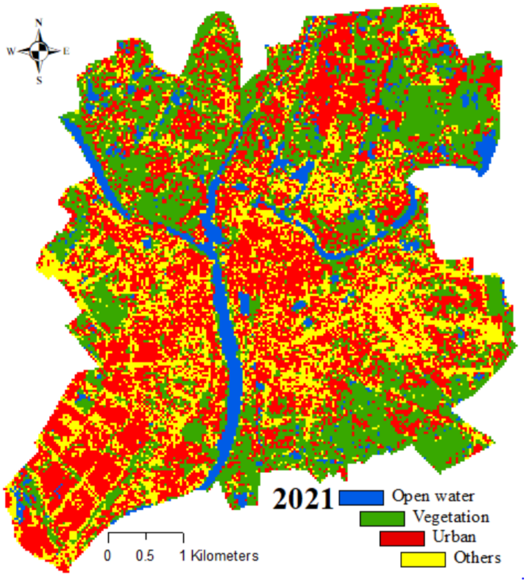


Figure 7. Random forest classification land cover map

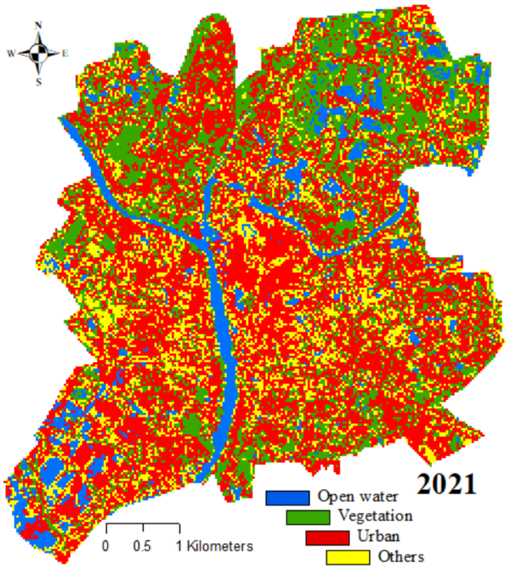
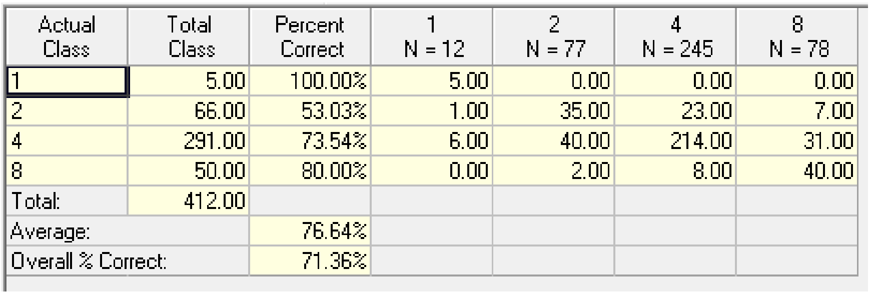


Figure 8. Maximum likelihood classification land cover map

The accuracy analysis was conducted to both processes with in 413 selected ground points using supporting materials. The overall accuracies of RF and ML method and are about 90% and 80% respectively providing that in this circumstance, RF method is more effective than the ML. According to land cover types classification, water was shown to obtain higher accuracy comparing to the other types. Figure 9 hightlighted a 100% correction in water extraction from data point. Obviously, the RF method acquires more varilables (indicators) in the process, and there are more restrict rules to train and assign samples into classes. However, there needs a further evaluation of this RF analysis at higher level classification.



Actual Class	Total Class	Percent Correct	1 N = 12	2 N = 77	4 N = 245	8 N = 78
1	5.00	100.00%	5.00	0.00	0.00	0.00
2	66.00	53.03%	1.00	35.00	23.00	7.00
4	291.00	73.54%	6.00	40.00	214.00	31.00
8	50.00	80.00%	0.00	2.00	8.00	40.00
Total:	412.00					
Average:		76.64%				
Overall % Correct:		71.36%				

Figure 9. The percentage of learning in evaluating the prediction success of each class, where water (class 1) has the highest percentage of 100 of accuracy

4. Conclusion

In overall, this research represented a potential approach of a machine learning method in image classification process for mapping land cover. Comparing to the common maximum likelihood, the random forests allowed to extract and generate information at a higher accuracy. They demonstrated robustness to noise and outliers, effectively handle high-dimensional datasets without overfitting, and exhibit high classification accuracy by capturing complex relationships in the data. The ability to assess variable importance helped identify influential features, while their capability to handle imbalanced data ensured accurate predictions for all classes. Additionally, random forests employed out-of-bag estimation for internal validation without the need for an additional validation dataset. However, they did have limitations, including limited interpretability due to their complex nature, computational intensity for large datasets, potential overfitting with noisy data, and challenges in extrapolation to unseen scenarios. Our future work will pay more attention on how to construct the next generation of decision trees and rules that satisfied a more detailed LULC map.

References

[1] L. J. J. Antonio Di Gregorio, Land Cover Classification System (LCCS) : Classification Concepts and User Manual, Rome, Italy: FAO, ISBN 92-5-104216-0, 2000.

[2] Y. Singh, "Significane of land use/landcover (LULC) maps," [Online]. Available: https://www.satpalda.com/blogs/significance-of-land-use-land-cover-lulc-maps.

[3] B. &. M. P. M. Tso, Classification methods for remotely sensed data (2nd ed.), CRC Press., 2009.

[4] M. A. M. D. K. H. J. C. Z. X. Y. M. D. S. A. H. W. C. E. G. S. S. A. C. A. B. A. G. F. &. S. C. Friedl, "Global land cover mapping from MODIS: algorithms and early results.," *Remote Sensing of Environment,* no. 83, pp. 287-302., 2002.

[5] G. M. Foody, " Field survey methods for reliable ground truth data collection for machine learning applications in land cover mapping," *nternational Journal of Remote Sensing,* vol. 20, no. 31, pp. 5331-5355, 2020.

[6] M. &. C. K. C. Herold, "The role of spatial metrics in the analysis and modeling of urban land use change," *Computers, Environment and Urban Systems,* vol. 3, no. 33, pp. 204-225., 2009.

[7] T. M. K. R. W. &. C. J. W. Lillesand, Remote Sensing and Image Interpretation (7th ed.)., Wiley, 2015.

[8] C. &. C. J. L. Glennie, " Airborne LiDAR for terrain and landcover mapping.," *In Environmental Remote Sensing, Elsevier.,* pp. 161-185, 2019.

[9] G. &. M. H. G. Vosselman, Airborne and terrestrial laser scanning, Whittles Publishing, 2010.

[10] T. Blaschke, "Object-based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 65, no. 1, pp. 2-16., 2010.

[11] Y. S. H. &. W. C. Zhang, "Data fusion in remote sensing: A review.," *ISPRS Journal of Photogrammetry and Remote Sensing,* no. 147, pp. 11-28, 2019.

[12] M. &. D. L. Belgiu, "Random forest in remote sensing: A review of applications and future directions," *SPRS Journal of Photogrammetry and Remote Sensing,* no. 14, pp. 24-31, 2016.

[13] M. &. M. P. M. Pal, "Support vector machines for classification in remote sensing," *nternational Journal of Remote Sensing,* vol. 5, no. 26, pp. 1007-2011, 2005.

[14] C. Chang, ntroduction to Geographic Information Systems (7th ed.), McGraw-Hill Education, 2013.

[15] K. C. &. W. C. E. Seto, "Monitoring Land Cover Change Using Remote Sensing Imagery," *Remote Sensing of Urban and Suburban Areas. Springer,* pp. 197-225, 2008.

[16] N. Horning, "Random Forests: An agorithm for image classification and generation continous fields datasets," *Computer Science,* 2010.

[17] M. Belgiu, "eo4geo.eu," 05 01 2022. [Online]. Available: http://www.eo4geo.eu/training/classification-random-forests/.

[18] L. F. B. G. Saheba Bhatnagar, "Drone Image Segmentation Using Machine and Deep Learning fo Mapping Raised Bog Vegetation Communities," *Remote Sensing,* p. 2602, 2020.

[19] C. R. L. Q. W. B. C. A. O. A. A. V. V. C. a. H. L. Tedros M. Berhane, "Decision - Tree, Rule-Based and Random Forest Classification of High Resolution Multispectra, Imagery for Wetland Mapping and Inventory," *Remote Sensing,* p. 580, 2018.