ACRS 2022
MONGOLIA

LAND COVER MAPPING IN CAMAU PROVINCE BY MACHINE LEARNING ALGORITHMS USING SENTINEL-2 IMAGERY

Tran Van Anh[1], Le Minh Hang[2], Tran Hong Hanh[1], Le Thanh Nghi[1], Tran Trung Anh[1], Nguyen Cam Chi[3], Ha Trung Khiên[4]

[1]Hanoi University of Mining and Geology, No.18 Vien Street, Duc Thang Ward, Bac Tu Liem, Hanoi, Vietnam,
Email: tranvananh@humg.edu.vn, tranhonghanh@humg.edu.vn, lethanhnghi@humg.edu.vn, trantrunganh@humg.edu.vn
[2]Le Quy Don Technical University, 236 Hoang Quoc Viet Street, Bac Tu Liem, Hanoi, Vietnam,
Email: leminhhang81@lqdtu.edu.vn
[3]Department of Survey, Mapping and Geographic Information Vietnam, 2 Đang Thuy Tram, Street, Bac Tu Liem, Hanoi, Vietnam,
Email: nguyencamchi@gmail.com
[4]Department of Surveying, National University of Civil Engineering, 55 - Giai Phong street - Dong Tam - Hai Ba Trung, Hanoi, Vietnam
Email: khienht@nuce.edu.vn

KEY WORDS: Xgboost, RF, SVM, Camau, GEE, Land cover

ABSTRACT: In this article, we have built a land cover map of Ca Mau province, Vietnam using 3 different classification methods random forest (RF), Support Vector Machine (SVM) and extreme gradient boosting (Xgboost). The study area is a mixed urban and rural area in the Mekong Delta, Vietnam with six land cover layers (LC). The satellite images used for classification are multi-temporal Sentinel-2 images from January to December 2021. The number of images in this period after cloud removal remains 17 images. The Median filtering method was used to generate an unique image in this time period for classification. The tool to do the classification is the Google Earth Engine platform. The sample is taken based on the land use map of Camau province in 2014 and Google Earth images. The number of samples taken for classification for all 3 methods is close to 4000 and the number of samples for the accuracy assessment is 3000 pixels. Overall (OA) error of SVM was 79.5%, kappa coefficient was 0.72 while the Xgboost method achieved 85.6% and Kappa: 0.79 and FR was OA: 86.5% and Kappa: 0.81. The image classified by RF method was selected to build the map at 1:50,000 scale.

1. INTRODUCTION

Ca Mau is a province located in the southernmost part of Vietnam with three sides bordering the sea. In recent years, due to climate change and sea level rise, subsidence and erosion of riverbanks and seabanks have occurred seriously. It is for this reason that the planning of agricultural land has also had to change accordingly and adapt to the situation of climate change. Currently, remote sensing technology has become popular and helps scientists in building various types of land use, land cover, planning or other fields easily. To be able to build these types of maps, the most important thing is the classification algorithm serving to separate different classes of land use data. In recent years machine learning methods have gained popularity and they have proven to be capable and accurate. One of the most widely used machine learning algorithms is random forest (RF) (Breiman 2001). The popularity of this algorithm is due to the fact that it can be used for both classification and regression purposes, and therefore can be used with categorical and continuous variables (Woznicki et al. 2019). Because of this versatility, RF has been used in a wide range of Earth sciences. Applications include forest modeling (Betts et al. 2017), land use (Araki, Shima and Yamamoto 2018), land cover (Nitze, Barrett and Cawkwell 2015), and object-oriented mapping (Kavzoglu 2017)).
RF comes from the traditional method of decision tree, by Rodriguez-Galiano et al. (2012), however with low number of trees, decision tree will cause large error, and thus RF is devised. The number of trees can be increased, and due to its combination of architectures in which several classification trees are trained on subsets of the training data, RF produces superior accuracy compared to classification trees ordinary type.
Support vector machine (SVM) is a machine learning algorithm based on statistical learning theory that has been widely used in the remote sensing community. They are seemed to perform better than most conventional classifiers (Huang et al 2002; Keuchelet et al 2003; Kavzoglu and Colkesen 2009; Su and Huang 2009). Furthermore, SVM even outperforms some novel pattern recognition methods, such as neural networks (Huang et al. 2002; Foody and Mathur 2004a, b). However, there are a number of parametric and non-parametric factors that can affect the performance of SVM, and it is necessary to investigate them so that SVM can be used with improved performance (Yang 2011).

Extreme gradient boosting (Xgboost) is a relatively new algorithm, first described by Chen and Guestrin (2016). One of the earliest remote sensing applications of Xgboost was made by (Georganos et al. 2018) which uses Bayesian parameter optimization on very high resolution WorldView-3 data. They found that Xgboost could outperform RF and SVM by 2–5% at larger sample sizes despite increasing computation time. (Man et al. 2018) compared five non-parametric classifiers using Landsat-8 data. They found that Xgboost outperformed SVM by 0.3%. Xgboost also was found to be slightly better than RF 0.2% in a recent six-layer LULC classification study using high resolution data from RapidEye (Hirayama et al. 2019).

These three machine learning methods all give good accuracy according to the articles surveyed above. However, the study area is a rather complicated area because there are many types of vegetation cover and seasonal changes. Therefore, the objective of this study is to compare the classification ability of three popular machine learning algorithms RF, SVM and XGBoost, to establish an landcover map for the study area of Ca Mau, Vietnam using Sentinel-2 images.

2. STUDY AREA AND MATERIALS

2.1. Study area

Ca Mau province is the end land of the country with 3 sides adjacent to the sea, geographical position is: East Sea borders with the East Sea with a coastline of 107 km. To the west and south borders the Gulf of Thailand with a coastline of 147 km. The North borders Bac Lieu province and Kien Giang province. Ca Mau is a lowland area, often flooded. Ca Mau has 5 main soil groups including: alkaline soil, peat soil, alluvial soil, saline soil and canal soil.

Ca Mau forest is a specific ecological type, mangrove coastal ecological forest is distributed along the coast with a length of 254 km. Besides, Ca Mau also has a melaleuca forest ecosystem located deep in the continent in the districts of U Minh, Tran Van Thoi and Thoi Binh with a scale of 35,000 ha. Mangrove area in Ca Mau accounts for 77% of mangroves in the Mekong Delta (Camau Web portal).

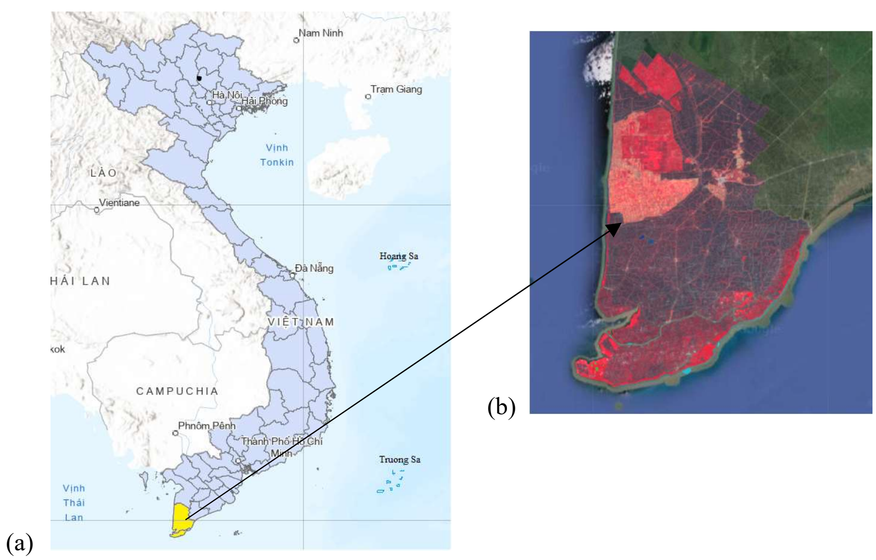

Figure 1. (a) Location of Camau in Vietnam; (b) The Pseudo color composite of Sentinel-2 image on Google Earth Engine

2.2. Materials

To establish a land cover map, we selected Sentinel-2 images. The supplied L2A product has been corrected for surface reflectance using the Dense Dark Vegetation (DDV) algorithm and the Atmospheric Pre-corrected Differential Absorption (APDA) algorithm (ESA Sentinel-2 User Handbook). This images has a spatial resolution of 10m and 20m with a repeat period of 12 days. The image acquiring period is from January 2021 to December 2021.

3. METHODOLOGY

3.1. Support vector machine

Support vector machine (SVM) was first described in (Cortes and Vapnik. 1995) based on the work of Vapnik (1982) and is a supervised learning technique commonly used in a wide range of remote sensing applications. The SVM algorithm finds the optimal minimization, i.e. decision boundary, of the outputs of the unambiguous classifier in a problem space. This decision boundary is called a hyperplane, and it separates the classification problem into a set of predefined classes that fit the training data. The algorithm goes through an iterative process to find the optimal hyperplane boundary in the n-dimensional classifier space to distinguish patterns in the training data, and then applies the same configuration to a set of separate assessment data. The dimensions in this context are the number of spectral bands, and the vectors are the individual pixels in a multiband composite (Mountrakis, Jungho, and Ogole 2011). There are different kernels through which the hyperplane boundary can be defined. A detailed mathematical description of this algorithm can be found in Cortes and Vapnik (1995).

3.2 Random Forest

Random Forest is an algorithm comprising of many single decision trees that act like unions. Each individual tree in Random Forest makes a prediction about the class and the class with the most votes becomes the prediction of that model (Figure 4). Random Forest model is very effective for image classification because it mobilizes hundreds of smaller models inside with different rules to make the final decision. Each sub-model can be different and weak, but according to the" wisdom of the crowd" principle, the classification result will be more accurate than using any single model.

As its name implies, Random Forest (RF) is based on: (1) Random means Randomness; (2) Forest equivalent to multiple decision trees.

The unit of RF is a decision tree algorithm with some hundred trees. Each decision tree is randomly generated from random sampling and only using a small part of random features from all variables in the data. Random forest makes on bagging. This means that at each split of the tree, the model considers only a small subset of features rather than all of the features of the model.

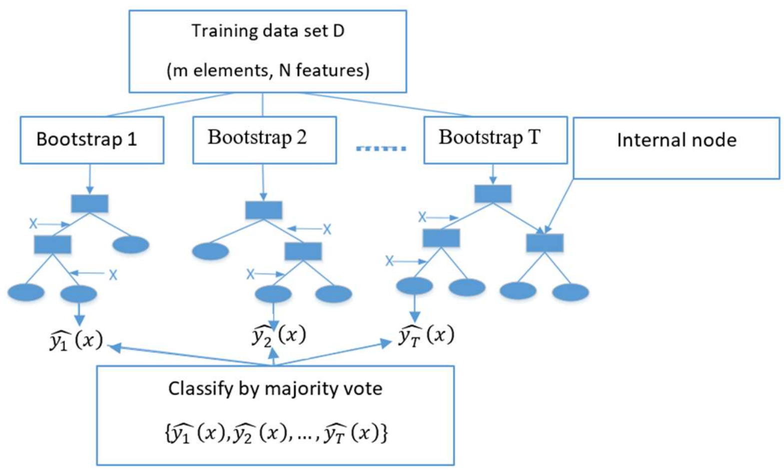

Figure2: Diagram of RF formation

3.3 Extreme Gradient Boosting

The basic idea of Gradient Boosting (GB) is to create a series of weak models that learn to complement each other. In other words, in Boosting, the following models will try to learn to limit the mistakes of the previous models.

The concept of gradient boosting involves basically three steps. First, a proper differentiable loss function should be identified that is suitable for the given problem. One benefit of the gradient boosting model is that for different loss functions, new algorithms are not required to be derived; it is enough that a suitable loss function be chosen and then incorporated with the gradient boosting framework. Second, a weak learner is created to make the predictions. In gradient boosting a decision tree is chosen as a weak learner. Specifically, regression trees are used that produces real value output for splits and whose output can be added together, allowing subsequent outputs of different models to be added. This approach enables the improvement of the residuals in the predictions leading to more precise predictions. The trees are created in a greedy manner and often certain constraints are imposed in order to ensure that the weak learners continue to be weak learners and still the trees can be created using a greedy approach. Third, creation of an additive model to add up the predictions of the weak learners so as to reduce the loss function. This process of adding the trees happens one at a time. The output produced in the new tree is then added to the output of the pre-existing sequence of trees in order to improve the final output of the model. This process stops once the proper optimized value for the loss function is reached.

Xgboost is a relatively new implementation of the GB that simultaneously optimizes the loss function while building the additive model (Chen and Guestrin 2016). The novelty of Xgboost lies in the fact that it comprises an objective function, which combines the loss function and a regularization term that controls model complexity. This enables parallel calculations and the maintenance of optimal computational speed. A detailed mathematical description of Xgboost is provided in Chen and Guestrin (2016)

3.4 Tool and steps

The tool used in this study is Google Earth Engine (GEE). GEE works through a JavaScript Application Online Interface (API) or Python called the Code Editor. On this interface, users can write and run scripts to share and repeat geospatial data processing and analysis processes. Code Editor helps users to perform all the functions available in Earth Engine. Input data can be directly exploited on data WEB pages without downloading to a computer, this is an advantage of GEE, allowing us to analyze data quickly without depending on the data and memory capacity of the computer (Tran et al, 2021). This is the interface of GEE (Figure 3).

To conduct image classification on GEE, it is necessary to go through some basic steps (Gorelick et al., 2017). Using Javascript programming language we followed the flow chart below to make three classification methods for Ca Mau area (Figure 4)

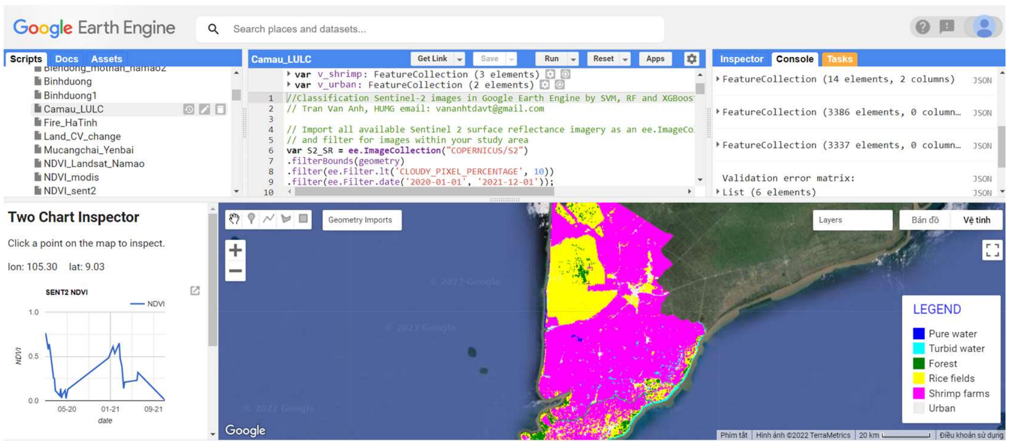

Figure 3. GEE's interface to write code for executing the commands

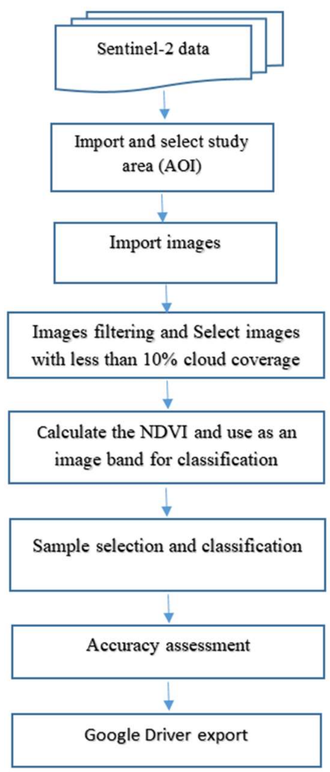

Figure 4. Image processing flowchart for image classification on GEE

3.4.1 Import and select the study area (AOI)

This is a limited area for image processing. The study area can be drawn directly on the screen or selected from an existing file. In the case of our research focusing on one province is Ca Mau, thus this area was created in shape file from ArcGIS software. This file was imported into GEE and used as a variable named “geometry”.

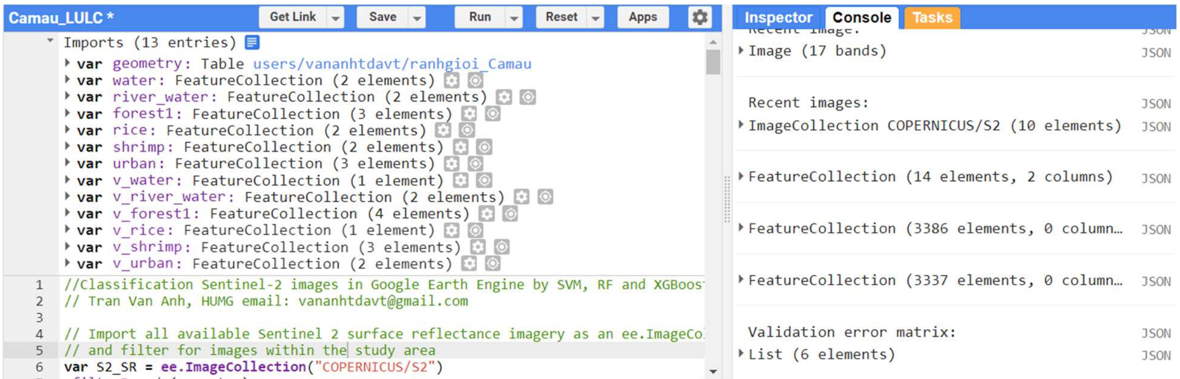

Figure 5. Import, select the study area and set time to select images on GEE

3.4.2 Filtering the images

Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission obtained from the website of the European space agency ESA - Copenicus. Importing images is shown in line 6, figure 4 .

Sentinel-2 images are filtered for the selection from January 2021 to December 2021. The set of images is then also filtered to remove those with more than 10% cloud coverage. Pixels that pass the requirements will be retained in the collection. A cloud mask is applied using a quality assessment band provided by (ESA, 2020) .

The composite images are made by calculation of the median values of each pixel (in all bands) in the sets. This step uses the function of GEE as ‘reducer’ to perform the compositing from the images in the set for generating a single output image. The cause of using this method for analyzing because using the median value reduces the effects of cloud (high values) and shadow (low values), and the resulting outputs are seamless mosaics appropriate for visualization purposes.

3.4.3 NDVI calculation

NDVI is calculated for Sentinel-2 time series images for the period January 2021 to December 2021. The total number of images in this period is 36. The selected images have less than 10% cloud coverage, so only 17 images satisfy the set conditions. The reason for determining the NDVI time series is that we want to determine the seasonal change of cropland so that we can separate agricultural land from forest land. NDVI after calculation will be updated into a band of images and will be used to classify images. The NDVI is a simple but effective and commonly used indicator for determining green coverage. The NDVI value ranges from −1 to +1. Values above 0.2 usually indicate the existence of green vegetation. Figure 6. NDVI code description on GEE.

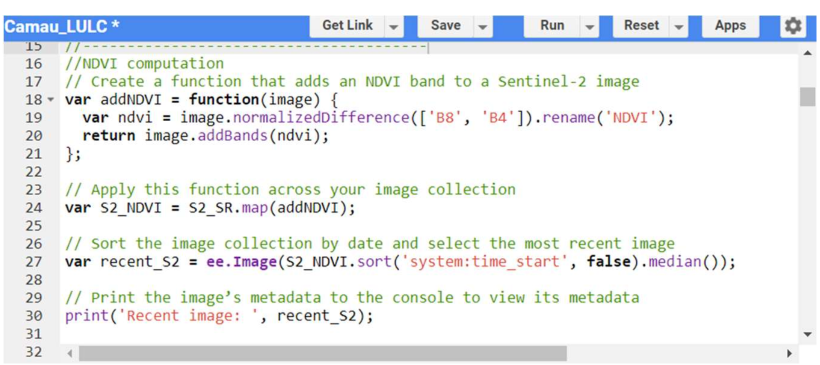

Figure 6. NDVI code description on GEE

3.4.4 Sample selection and classification

After the series of images were selected, the median for the set of images has been calculated which is to remove the effect of noise and clouds on the image. Sampling was taken on the land use map on 1:50000 scale in 2014 and google Earth image. The image bands used for the classification are Band 1 to Band 8 and NDVI. We selected 3 classification methods: SVM, RF and XGBoost. All three methods can use the same sample for classifying or to assess accuracy, therefore this set of samples was taken together. Nearly 3386 pixels were sampled for 6 different layers and 3000 pixels were sampled to evaluate the accuracy for the three classifiers.
Since two machine learning methods RF and XGBoost use a decision tree platform, the number of trees must be selected. After testing we found that the number of trees is 300 is satisfactory. Figure 7. shows an example of the code used to classify by RF method with a tree number of 300.

Camau_LULC * | Get Link | Save | Run | Reset | Apps

```
//-----------------------------------------------------
// Classification for Sentinel-2 image

// Gộp các vùng mẫu (merge feature collection)
var classNames = water.merge(river_water).merge(forest1).merge(rice).merge(
print(classNames);
// create training area
var S2_NDVI_med = S2_NDVI.median().clip(geometry);
var bands = ['B2', 'B3', 'B4','B5', 'B6', 'B7', 'B8', 'NDVI'];
var training = S2_NDVI_med.select(bands).sampleRegions({
  collection: classNames,
  properties: ['landcover'],
  scale: 30
});
print(training);
// Classification
var classifier = ee.Classifier.smileRandomForest(300).train({
  features: training,
  classProperty: 'landcover',
  inputProperties: bands
});
//Run the classification
var classified = S2_NDVI_med.select(bands).classify(classifier);
//Display classification
Map.addLayer(classified,
{min: 0, max: 5, palette: ['#0000FF', '#00FFFF', '#008000', '#FFFF00', '#FF
'LULC');
```

Figure 7. Classification code description on GEE

3.4.5 Accuracy assessment

The accuracy of the results is based on confusion matrix, overall accuracy (OA) and Kappa and Consumer's accuracy, Producer's accuracy (Aronoff., 1982). The command lines to calculate the accuracy are shown in the Figure 8.

Camau_LULC | Get Link | Save | Run | Reset | Apps

```

//------------------------------------------
// Validation
//Merge into one FeatureCollection
var valNames = v_water.merge(v_river_water).merge(v_forest1).merge(v_rice).merg
//Sample your classification results to your new validation areas
var validation = classified.sampleRegions({
  collection: valNames,
  properties: ['landcover'],
  scale: 30,
});
print(validation);
//Compare the landcover of your validation data against the classification resu
var testAccuracy = validation.errorMatrix('landcover', 'classification');
//Print the error matrix to the console
print('Validation error matrix: ', testAccuracy);
//Print the overall accuracy to the console
print('Validation overall accuracy: ', testAccuracy.accuracy());

// Calculate consumer's accuracy, also known as user's accuracy or
// specificity and the complement of commission error (1 - commission error).
print("Consumer's accuracy", testAccuracy.consumersAccuracy());

// Calculate producer's accuracy, also known as sensitivity and the
// compliment of omission error (1 - omission error).
print("Producer's accuracy", testAccuracy.producersAccuracy());

// Calculate kappa statistic.
print('Kappa statistic', testAccuracy.kappa());
```

Figure 8. Accuracy assessment om GEE

4. RESULTS AND DISCUSSION

According to the flow chart in Figure 4, the classification results of the Ca Mau area from the three methods are shown in Figure 9. Tables 1,2,3 show the results of the accuracy assessment.

The results show that with the three selected classification methods, RF and XGBoost both give better accuracy than SVM. Table 1, 2,3 show the results of the accuracy assessment including the error matrix, the overall error (OA), the Consumer's accuracy and the Producer's accuracy. With the forest layer, both RF and XGBoost methods have higher classification accuracy. With the error matrix, the ability to separate the Forest with the Rice fields layer is much worse when up to 433 pixels are mixed into the rice fields layer while with the other two methods it is only 206 pixels with Xgboost and 199 pixels with RF. According to shrimp farms classifications all three methods have similar errors. The study area is low land with many shrimp farms and that is the reason why it is difficult to distinguish between shrimp fields and wet rice fields. Therefore, it is easy to see that with all these three classifications, the confusion between shrimp fields and wet rice fields is quite high. Therefore we have added the NDVI value to classify along with the image bands for increasing accuracy. Similar to shrimp farms, this area has not much urban land and is mainly mixed with cropland, so the classification is also confused with other layers. The UA value of urban is less than 40% for all three methods.

Overall (OA) error of SVM was 79.5%, kappa coefficient was 0.72 while the Xgboost method achieved 85.6% and Kappa: 0.79 and FR was OA: 86.5% and Kappa: 0.81.

Through the above evaluations, the classification results from RF were selected as having the highest Kappa and Kappa. Combined with the field investigation, we have established the land cover map of Ca Mau province on ArcGIS 10.4 software (Figure 10).

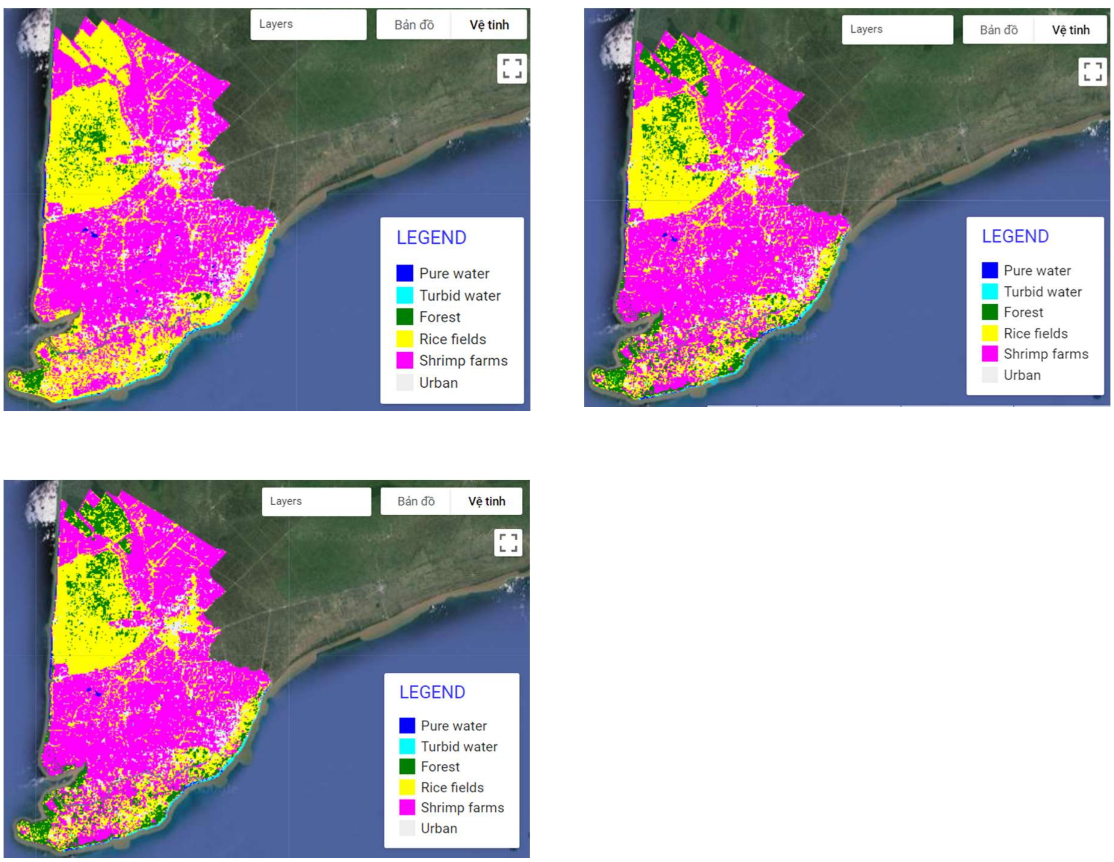

Figure 9. (a) SVM Classification, (b) RF Classification, (c) XGBoost Classification

Tab. 1. Accuracy of XGBoost classification with 300 trees.

Accuracy				**Confusion matrix**					
Classes	UA (%)	PA (%)		Pure water	Turbid water	Forest	Rice fields	Shrimp farms	Urban
Pure water	88.3	96.8	Pure water	651	0	0	0	21	0
Turbid water	98.6	83.1	Turbid water	67	375	0	0	7	2
Forest	98.9	86.6	Forest	0	0	1542	206	1	41
Rice fields	45.3	97.6	Rice fields	0	0	6	249	0	0
Shrimp farms	81.4	49.3	Shrimp farms	19	5	11	94	180	56
Urban	36.7	75.0	Urban	0	0	0	0	12	36
	Over all accuracy: 85.6% and Kappa: 0.79								

Tab. 2. Accuracy of RF classification with 300 trees.

Accuracy				**Confusion matrix**					
Classes	UA (%)	PA (%)		Pure water	Turbid water	Forest	Rice fields	Shrimp farms	Urban
Pure water	91.3	95.6	Pure water	643	0	0	0	29	0
Turbid water	99.7	88.4	Turbid water	43	399	0	1	7	1
Forest	99.8	88.8	Forest	0	0	1558	190	1	4
Rice fields	45.2	99.2	Rice fields	0	0	2	253	0	0
Shrimp farms	78.4	48.7	Shrimp farms	18	1	0	115	178	53
Urban	38.3	75.0	Urban	0	0	0	0	12	36
	Over all accuracy: 86.5% and Kappa: 0.81								

Tab. 3. Accuracy of SVM classification

Accuracy				**Confusion matrix**					
Classes	UA (%)	PA (%)		Pure water	Turbid water	Forest	Rice fields	Shrimp farms	Urban
Pure water	89.6	95.2	Pure water	640	0	0	0	32	0
Turbid water	99.7	88.9	Turbid water	43	401	0	0	2	5
Forest	99.9	75.2	Forest	0	0	1319	433	0	1
Rice fields	31.7	99.6	Rice fields	0	0	1	254	0	0
Shrimp farms	80.9	47.6	Shrimp farms	22	1	0	112	174	56
Urban	34.0	66.6	Urban	9	0	0	0	7	32
	Over all accuracy: 79.5% and Kappa: 0.72								

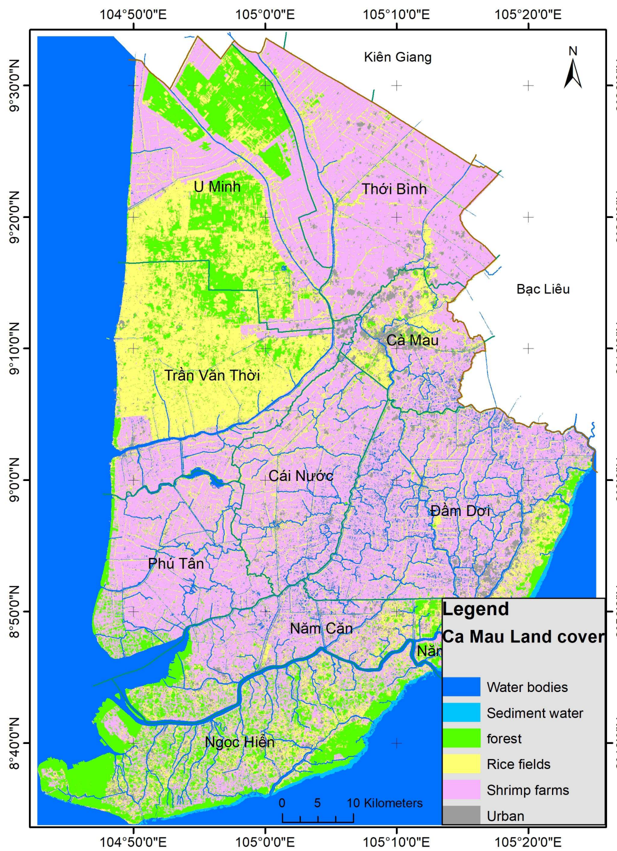

Figure 10. Land cover map of Ca Mau province

5. CONCLUSION

By using Sentinel-2 satellite images to build the land cover map in the area of Ca Mau province, Vietnam, we found that Sentinel-2 images have quite good resolution to be able to build a land cover map with scale of 1:50,000. Besides, the 2014 Ca Mau land use map was used to sample and evaluate the accuracy for the classifications

Among three image classification methods, SVM, XGBoost and RF, the XGBoost and RF methods give almost the same accuracy but the RF gives slightly better accuracy. SVM is the method with the lowest accuracy of the three methods and especially with this method, the distinction between two similar objects, forest and rice fields, is much worse than RF and XGboost methods.

The tool used toconduct image classification is Google Earth Engine. This is a pretty convenient cloud computing platform. This platform has made it quick and convenient to experiment with image classification methods.

ACKNOWLEDGEMENTS

This work is supported and funded by the Scientific Research project of Vietnam Ministry of Education and Training, code: B2022-MDA-13. In addition, the authors would like to thank Google provided Google Earth Engine development platform for processing data.

REFERENCES

Araki, S., M. Shima, and K. Yamamoto., 2018, Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan: Science of The Total Environment, v. 634, p. 1269-1277.

Betts, M. G., C. Wolf, W. J. Ripple, B. Phalan, K. A. Millers, A. Duarte, S. H. Butchart, and T. Levi., 2017, Global forest loss disproportionately erodes biodiversity in intact landscapes: Nature, v. 547, p. 441-444.

Breiman, L., 2001, Random forests: Machine learning, v. 45, p. 5-32.

Chen, T., and C. Guestrin., 2016, Xgboost: A scalable tree boosting system: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, p. 785-794.

Foody, G. M., and A. Mathur., 2004, Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification: Remote Sensing of Environment, v. 93, p. 107-117.

Georganos, S., T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, and E. Wolff., 2018, Very high resolution object-based land use–land cover urban classification using extreme gradient boosting: IEEE geoscience and remote sensing letters, v. 15, p. 607-611.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore., 2017"Google Earth Engine: Planetary-scale geospatial analysis for everyone." Remote Sensing of Environment 202: p18-27.

Hirayama, H., R. C. Sharma, M. Tomita, and K. Hara., 2019, Evaluating multiple classifier system for the reduction of salt-and-pepper noise in the classification of very-high-resolution satellite images: International journal of remote sensing, v. 40, p. 2542-2557.

Huang, C., L. Davis, and J. Townshend., 2002, An assessment of support vector machines for land cover classification: International Journal of remote sensing, v. 23, p. 725-749.

Kavzoglu, T., 2017, Object-oriented random forest for high resolution land cover mapping using quickbird-2 imagery, Handbook of neural computation, Elsevier, p. 607-619.

Kavzoglu, T., and I. Colkesen., 2009, A kernel functions analysis for support vector machines for land cover classification: International Journal of Applied Earth Observation and Geoinformation, v. 11, p. 352-359.

Lin, Y., F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang., 2011, Large-scale image classification: fast feature extraction and svm training: CVPR 2011, p. 1689-1696.

Man, C. D., T. T. Nguyen, H. Q. Bui, K. Lasko, and T. N. T. Nguyen., 2018, Improvement of land-cover classification over frequently cloud-covered areas using Landsat 8 time-series composites and an ensemble of supervised classifiers: International Journal of Remote Sensing, v. 39, p. 1243-1255.

Mountrakis, G., J. Im, and C. Ogole., 2011, Support vector machines in remote sensing: A review: ISPRS Journal of Photogrammetry and Remote Sensing, v. 66, p. 247-259.

Nitze, I., B. Barrett, and F. Cawkwell, 2015, Temporal optimisation of image acquisition for land cover classification with Random Forest and MODIS time-series: International Journal of Applied Earth Observation and Geoinformation, v. 34, p. 136-146.

Sentinel E, User Handbook., 2015. ESA Standard Document, 64p

Su, L., and Y. Huang., 2009, Support vector machine (svm) classification: comparison of linkage techniques using a clustering-based method for training data selection: GIScience & Remote Sensing, v. 46, p. 411-423.

Tran, V. A., T. L. Le, N. H. Nguyen, T. N. Le, and H. H. Tran., 2021, Monitoring Vegetation Cover Changes by Sentinel-1 Radar Images Using Random Forest Classification Method: Inżynieria Mineralna.

Vapnik, V.. 1982. Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (springer Series in Statistics). New York, NY: Springer-Verlag

Veropoulos, K., C. Campbell, and N. Cristianini., 1999, Controlling the sensitivity of support vector machines: Proceedings of the international joint conference on AI, p. 60.

Woznicki, S. A., J. Baynes, S. Panlasigui, M. Mehaffey, and A. Neale., 2019, Development of a spatially complete floodplain map of the conterminous United States using random forest: Science of the total environment, v. 647, p. 942-953.