A STUDY ON OIL PALM CLASSIFICATION FOR RANONG PROVINCE USING DATA FUSION AND MACHINE LEARNING ALGORITHMS

Morakot WORACHAIRUNGREUNG [1], ***Kunyaphat THANAKUNWUTTHIROT*** [2], ***Nayot KULPANICH*** [1]

DOI: 10.21163/GT_ 2023.181.12

ABSTRACT:

Oil palm is a vital force in driving the energy business. In 2020, Thailand had 9,954.27 sq.km. (around 6,220,799 Rai) of oil palm plantations, ranking third in the world after Indonesia and Malaysia. Ranong has the highest oil palm crop yield per Rai in Thailand. Notwithstanding, it is challenging to classify land use accurately and keep it up to date by using only labor, due to the need for a number of laborers and high labor costs. Moreover, land use/land cover cannot use spectral information classification alone. Nevertheless, machine learning is a popular data estimation technique that enables a system to learn from sample data; however, there are few studies on its use for data fusion techniques in order to classify land use/land cover, especially concerning oil palm. Therefore, we aim to apply machine learning and data fusion to classify land use/land cover, especially for oil palm. After a multicollinearity test of spectral information and ancillary variables, Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1, NDWI, NDVI and LST were selected with a threshold of correlation coefficients. A stepwise stack of six inputs was created. The first stack included only Surface Reflectance (SR) of Blue, Near Infrared and SWIR-1. NDWI, NDVI and LST were added later. ID4 (Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1, NDWI, NDVI and LST) in the random forest model resulted in OA being 0.9341 and KC being 0.9239, which was the highest among 12 models. ID4 in the random forest model provided the classification results for oil palm very close to the factual number per the figure of 2.90 sq.km (around 1,814 Rai) from the Department of Land.

Key-words: *Oil palm, Ranong, Data fusion, Machine learning, Remote Sensing*

1. INTRODUCTION

All industries inevitably need energy to drive their industries. Oil palm is a crucial factor in driving the energy business. From the 1970s to the 2020s, oil palm area has dramatically doubled, and such an increase in oil palm plantations affects the ecosystem. At present, alternative energy has been used to replace fossil fuels, which includes non-renewable petroleum, natural gas and coal as the main sources of electricity and energy used in daily life. Biomass, or biological energy, obtained from palm oil is an alternative as a renewable energy source; for instance, biodiesel renewable energy can replace fuel for future transportation and can also be used as a raw material in soap and foods such as condensed milk, ice cream and butter. This is in line with 17 sustainable development goals as presented by the United Nations (Shaharum N. S., et al., 2020).

In 2020, Thailand had 9,954.27 sq.km. (6,220,799 Rai) of oil palm plantations, which ranked third in the world after Indonesia and Malaysia. The southern region of Thailand has 8,511.36 sq.km. (5,319,602 Rai) of oil palm plantations, and has a total area of 8,077.63 sq.km. (5,048,519 Rai) that can currently produce actual production volume. The province with the largest growing area in the southern region is Krabi, with a total area of 1,873.30 sq.km. (1,170,815 Rai) of oil palm plantation, but the province with the highest crop yield per Rai is Ranong, serving 2,980 kilograms per 1,600 sq.m. (Land Development Department, 2021). Accurate and up-to-date data is important for land management. Notwithstanding, it is quite difficult to classify land use accurately and up-to-date using labor, due to the need of the number of laborers and high labor costs. Using satellite images to classify

[1] *Suan Sunandha Rajabhat University, Faculty of Humanities and Social Sciences, Geography and Geo-Informatics Program, 1 U-Thong nok Road, Dusit, Bangkok 10300 Thailand, Morakot.wo@ssru.ac.th,nayot.ku@ssru.ac.th*

[2] *Suan Sunandha Rajabhat University, Faculty of Fine and Applied Arts, Digital Design and Innovation Program, 1 U-Thong nok Road, Dusit, Bangkok 10300 Thailand, kunyaphat.th@ssru.ac.th*

land use can well solve the problems of the number of laborers and high labor costs. Until the present, satellite images have been used to classify land use in numerous cases (George, Padalia & Kushwaha, 2014).

At present, satellite imagery is applied in the classification of many field crops. The use of satellite images to pinpoint old palm areas is widely popular (Srestasathiern & Rakwatin, 2014; Li, Dong , Fu, & Yu, 2019). To illustrate, Thenkabail used satellite images at 4-meter resolution captured from an IKONOS satellite to study oil palm biomass (Thenkabail, et al., 2004). Gutiérrez used satellite images at 250-meter resolution from MODIS to study oil palm area spanning 939,204 square kilometres (Gutiérrez-Vélez & DeFries, 2013). J. Miettinen studied oil palm plantations across Southeast Asia and Peninsular Malaysia, Sumatra, Java, Borneo, Sulawesi and Mindanao islands (Miettinen, Shi, Tan, & Liew, 2012). This study classifies land use by separating data into 13 layers, including mangrove forests, forests, rain forests and oil palm. By using high-resolution data, it can also help classify oil palm areas. In 2011, Shafri used the maximum likelihood classifier to classify Ganoderma disease infected oil palms with an accuracy of 82 percent (Shafri, Hamdan, & Saripan, 2011). Moreover, Kulpanich et al. used the UAV images to collect relevant data to forecast oil palm yields (Kulpanich, et al., 2022). However, the limitation of UAV images is that the large area will take a lot of time and budget for the operation. From the above statement, it was found that moderate-resolution satellites (MODIS) can classify oil palms, so it is believed that LANDSAT9, with a higher spatial resolution than MODIS, will also be able to classify oil palms.

Regarding the classification algorithm, Morel successfully differentiated forest from oil palm using k-means and an MLC algorithm (Morel, Fisher, & Malhi, 2012). In addition, Cheng successfully classified land cover using LANDSAT and ALOS-PALSAR through SVM and Minimum Distance algorithms (Cheng, Yu, Cracknell, & Gong, 2016). The study found that, for the classification of two areas, SVM was better than Minimum Distance algorithms that give satellite images from LANDSAT and ALOS-PALSAR. Furthermore, Cheng’s study covered the areas of Malaysia, Indonesia, Thailand and Nigeria, with an accuracy of over 94 percent for those aforementioned countries (Cheng, et al., 2018). But it found little application using machine learning in classifying oil palms by satellite imagery.

Nowadays, machine learning in classification has been widely adopted (Worachairungreung, et al., 2021; Worachairungreung, Thanakunwutthirot, & Ninsawat, 2019) Machine learning is mostly applied on oil palm classification for interpretation. Nooni used Support Vector Machine learning models to classify oil palm areas (Nooni, et al., 2014). Sitthi used Naive Bayes classifiers to identify what is covered in given areas, (Sitthi, et al., 2016) and Mubin used a convolutional neural network as a deep learning method to identify young and mature oil palm trees) Mubin, et al., 2019). Nevertheless, it is rare to find classification research comparing multiple algorithms given two types of satellite images or more.

The classification of LULC is complex. Currently, data fusion techniques are used to help classify many LULCs. Data Fusion is a method or tool to combine remote sensing data from different sources and multiply them to create new data in order to obtain representative data. Some researchers use a Digital Elevation Model (DEM) and SAR derived features that contribute the most to building damage classification. Classification results showed an overall accuracy of >90% and an average of >67% (Adriano, et al., 2019). Some researchers used a data fusion method and NDVI time-series analysis-based phenology extraction. The Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM) technique accurately blended SPOT5 and MODIS NDVI in Shandong Province, China, where counties tested phenology detecting methods with data fusion techniques (Yin, et al. 2019). Some researchers have used mono-temporal and multi-temporal LULC classifications and auxiliary data to determine LULCC in southwest Burkina Faso's varied landscape. Multi-temporal classification outperformed mono-temporal classification in the research area (Zoungrana , et al., 2015). According to the study, data fusion improves classification outcomes. Therefore, this study wanted to use such a technique to classify oil palm. Oil palm is an economically important plant in southern Thailand, but data fusion and machine learning are rarely used to classify oil palms, so in this study, Landsat 9 satellite imagery with 30 meter spatial resolution was used to classify it. In addition, data fusion and machine learning algorithms such as SVM, Random Forest and CART were used to classify palm, but this study studied only the Ranong dataset. Finally, it is hoped that this study will help classify oil palm in Thailand, as well as other countries in the region.

2. STUDY AREA

Ranong Province is in the southern region of Thailand, with an area of 3,426 sq.km. (2,141,250 Rai). The province is comprised of five districts: Mueang District, La-un-District, Kapoe District, Kra Buri District, and Suk Samran District, with an elevation range between 0-1,388 meters above sea level. Ranong Province is located in the southwest part of Thailand, and entirely influenced by the southwest monsoon. It receives more abundant rainfall than other provinces and it falls most of the year. Most areas of the province are covered by rubber, orchard, forest, mangrove and oil palm. (**Fig. 1**).

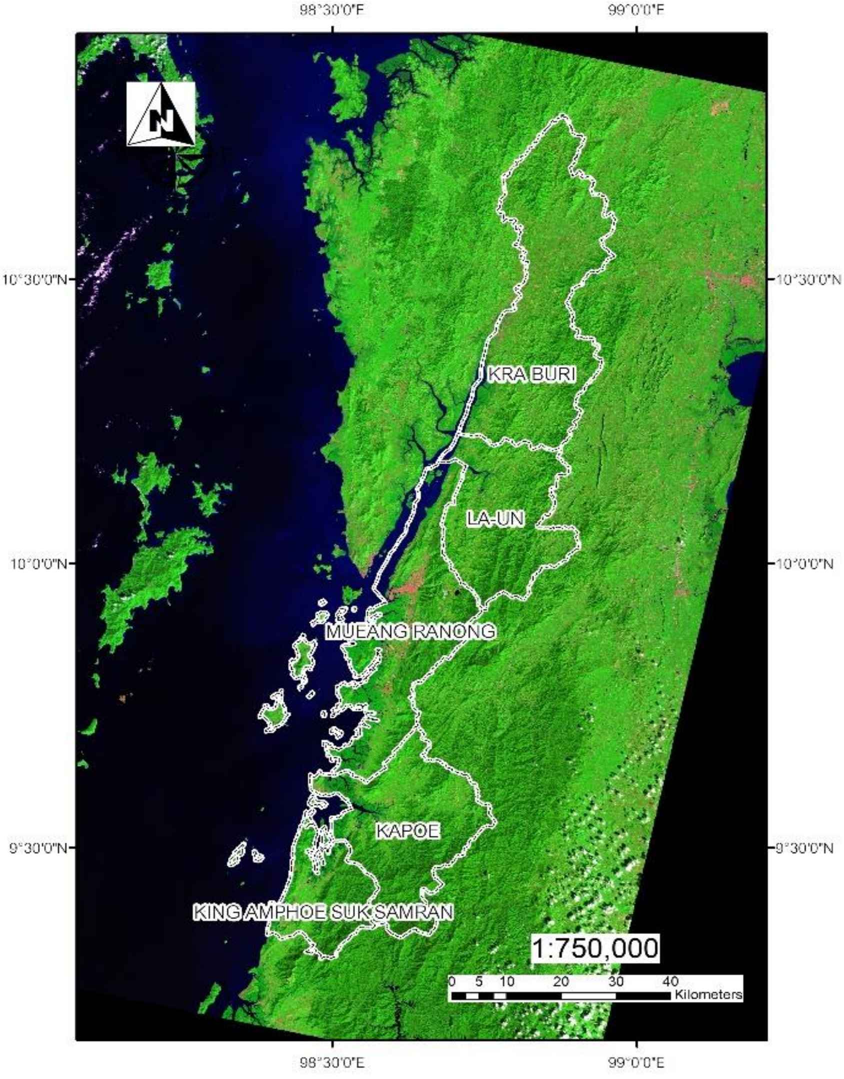

Fig. 1. Ranong Province.

3. DATA AND METHODS

In this study, satellite imagery and data were used. We then take satellite imagery analyzing the land surface, temperature, soil index, water index and vegetation index, and run the analysis results through data fusion. We then use a multicollinearity method to reduce data redundancy. We apply machine learning algorithms by dividing the training and testing datasets into 80:20 proportions. Finally, we compare the results. **Figure 2** shows the overall methodology mentioned above.

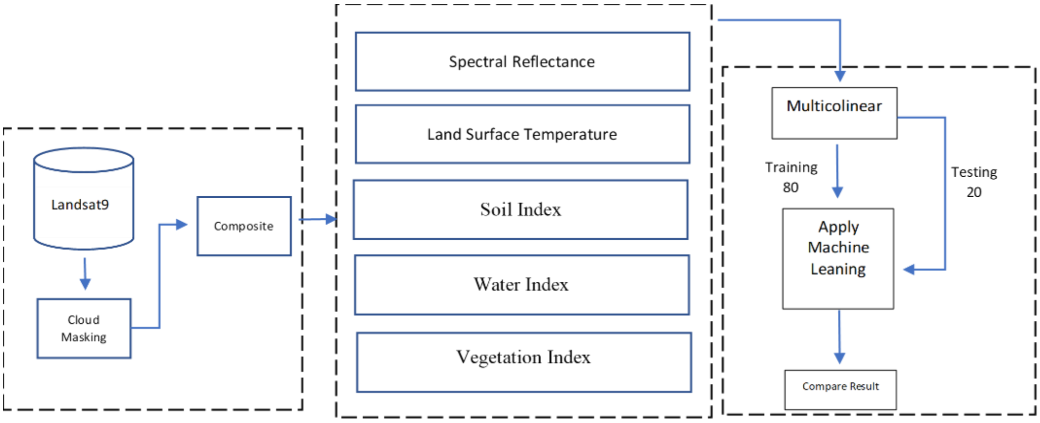

Fig. 2. Overall Methodology

3.1. Landsat 9 Spectral Reflectance Data

In this study, Landsat 9 was selected because the Landsat program continues its mission to capture repetitive observations worldwide for monitoring, comprehending and managing Earth's natural resources, with Landsat 9 under collaboration between the U.S. Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). Researchers rely on the USGS's Landsat archival data, which has been freely available since 1972, to map changes to the land's surface, but it is required to be pre-processed to make it usable. The researchers chose and altered remotely sensed products using an on-demand interface provided by the USGS Earth Resources Observation and Science (EROS) Center. **Table 1** shows the Landsat 9 details.

Table 1.

Landsat 9 Spectral Reflectance Data

Spectral	Wavelength in micrometers	Resolution in meters
Operational Land Imager		
Band 1—Ultra blue (coastal/aerosol)	0.435–0.451	30
Band 2—Blue	0.452–0.512	30
Band 3—Green	0.533–0.590	30
Band 4—Red	0.636–0.673	30
Band 5—Near infrared (NIR)	0.851–0.879	30
Band 6—Shortwave infrared (SWIR) 1	1.566–1.651	30
Band 7—Shortwave infrared (SWIR) 2	2.107–2.294	30
Band 8—Panchromatic	0.503–0.676	15
Band 9—Cirrus	1.363–1.384	30
Thermal Infrared Sensor		
Band 10—Thermal infrared (TIR) 1	10.60–11.19	100
Band 11—Thermal infrared (TIR) 2	11.50–12.51	100

The Landsat 9 dataset contains surface reflectance from an Operational Land Imager, top of atmospheric (TOA) reflectance, and TOA brightness for temperature in Kelvin, as well as spectral indices, including a Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), and Normalized Difference Moisture Index (NDMI).

The Landsat 9 scene of path 130 and row 53 was projected using UTM with datum WGS84, and was acquired during a field survey (January 2022). This study adopted SR Bands TOA brightness temperature, Normalized Difference Vegetation, Normalized Difference Vegetation Index, Normalized Difference Water Index, and Soil Index at 30-meter resolution. The researchers only considered the optical bands (2 to 7) for classification among all SR bands. The TOA brightness temperature (only band 10) was utilized to estimate land surface temperature.

3.2. Field Data

Table 2 and **figure 3** show field data. Field data from 1,600 field data points was obtained from the Global Positioning System (GPS). The field data consisted of rubber, oil palm, orchard, forest 1, forest 2, mangrove, built-up area, and water bodies. The data were divided into 2 sets: set 1 was for modeling, and set 2 was for testing accuracy of the model under the proportion of 80:20, respectively.

Table 2.
Field Data

Land Cover Class	NO. Field Data
Bare Soil	36
Rubber	51
Orchard	92
Forest	160
Evergreen Forest	390
Mangrove	245
Oil Palm	420
Built up Area	104
Water Bodies	104

3.3. Auxiliary Variables

In this study, Landsat 9 consisted of SR bands 2-7 and auxiliary data of NDVI, SI and NDWI, including Land Surface Temperature estimation. The purpose of using such auxiliary variables was to improve classification accuracy.

3.3.1. Land Surface Temperature

There are many ways to calculate Land Surface Temperature (LST). Previous research comparing different LST estimation methods favored the Radiative Transfer Equation (RTE)-based technique and band 10 over band 11. (Zhou, et al., 2012; Jiménez-Muñoz, et al., 2014; Santos, et al., 2018; Rehman, et al., 2021). In this study, we used Landsat 9 TIRS Band 10 to estimate LST by using a technique based on the Radiative Transfer Equation (RTE), as described in Equation (1).

$$\mathrm{LST}_{B10} = \tau i(\theta)\mathrm{EiBi}\ (Ts) + \left((1 - Ei)I^{\downarrow}\right) + I^{\uparrow} \tag{1}$$

i(θ): atmospheric transmission for band 10 when view zenith angle is θ;
Ei: surface emissivity of the band 10;
Bi(Ts) : a ground radiance;
I^↓ : Downwelling path radiance;
I^↑ : Upwelling path radiance.

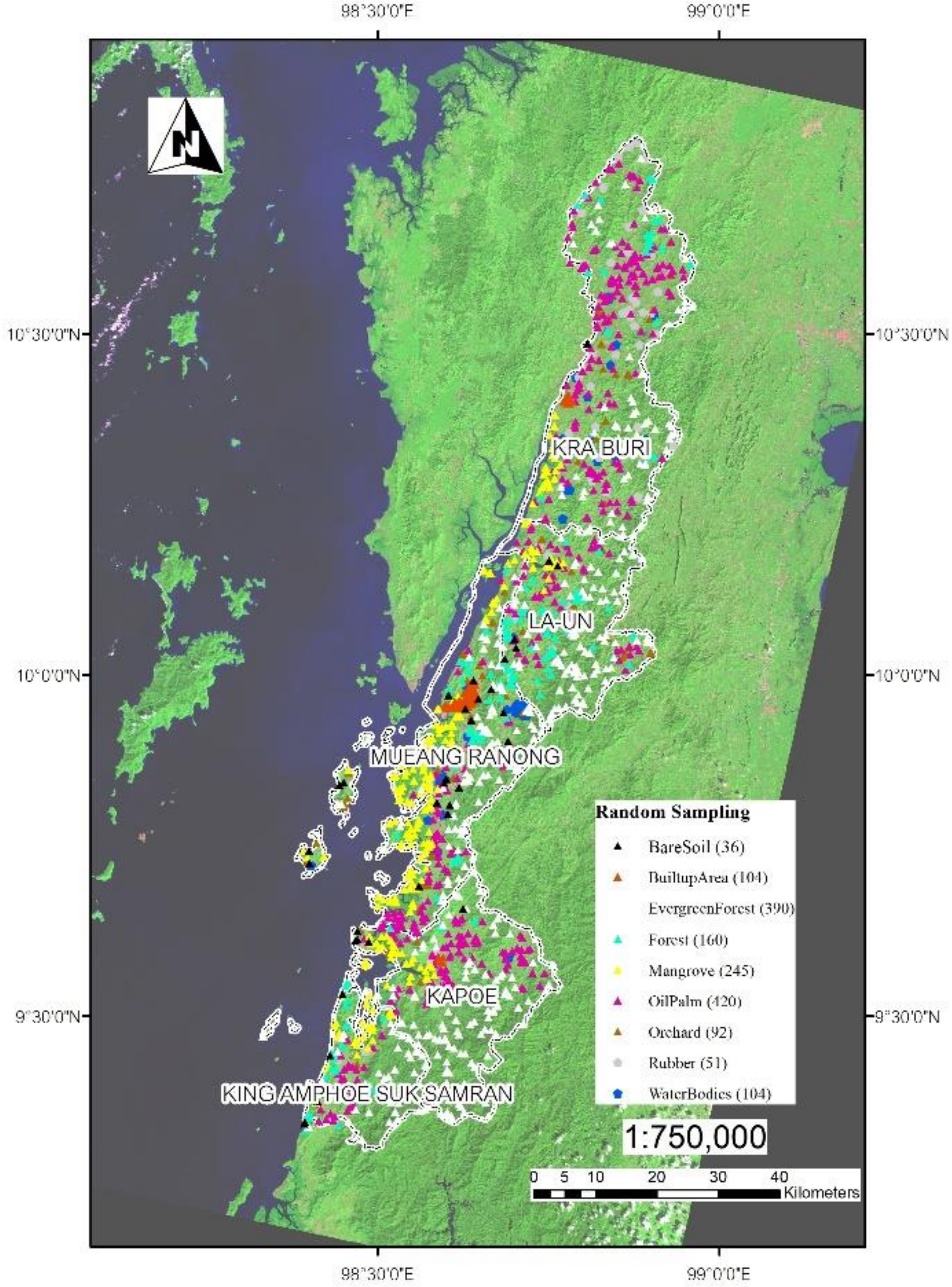

Fig. 3. Field Data.

According to Plank’s law, ground radiance Bi(Ts) can be expressed as:

$$Bi(Ts) = \frac{2hc^2}{\lambda_{Bi}^5\left(\exp\left(\frac{hc}{\lambda_{Bi}kTs}\right) - 1\right)} \tag{2}$$

where c is the speed of light (c = 2.9979 x 108 m/s), h is the Planck constant (h = 6.6261 × 10-34 J.s), k is the Boltzmann constant (k = 1.3806 ×10-23 J/K), λ represents the wavelength of TIRS bands (B10 = 10.602), and Ts is TOA brightness temperature.

3.3.2. Soil Index

Soil is a substance with several chemical and physical components (Thenkabail, et al., 2004). In this study, we used the Normalized Difference Soil Index (SI) proposed by Deng (Deng, et al., 2015). They studied the spectrum reflectance of soil samples using Landsat-5 data and discovered that the mean reflectance values of bands NIR, SWIR1 and SWIR2 are greater than those of visible bands. In addition, they examined all potential band normalized difference combinations and concluded that the index obtained from the SWIR2 and green bands are superior for mapping soil information.

$$SI = \frac{Green - SWIR2}{Green + SWIR2} \tag{3}$$

3.3.3. Water Index

McFeeters came up with the NDWI, which is the ratio of the green band to the NIR band (McFEETERS, 1996). Xu revised the NDWI to become the ratio of the green band to the SWIR band (Xu H. , 2006). Using Landsat 8 data, a previous study compared both methods developed by McFeeters and Xu, and found that the best way to map water bodies is to use both Green and SWIR (Du, et al., 2014). Hence, in this study, we applied the equation for the water index made by Xu. Green Bands and SWIR1 were used in this method.

$$NDWI = \frac{Green - SWIR1}{Green + SWIR1} \tag{4}$$

3.3.4. Vegetation Index

The normalized difference vegetation index (NDVI) is one of the most often used vegetation indices, and its value in satellite assessment and monitoring of global plant coverage has been well recognized over the last two decades (Huete, Justice, & Liu, 1994; Leprieur, et al., 2000). The formula is described as follows.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{5}$$

3.3.5. Multicollinearity Test

Implementing land use classification requires testing for SR and ancillary data redundancy. The solution to such a problem in this study was to use a correlation matrix between the SR bands/ancillary variables. From the study, if the correlation value was greater than 0.7, it was necessary to reduce several redundant variables before importing to a machine learning model, as shown in **figure 4**.

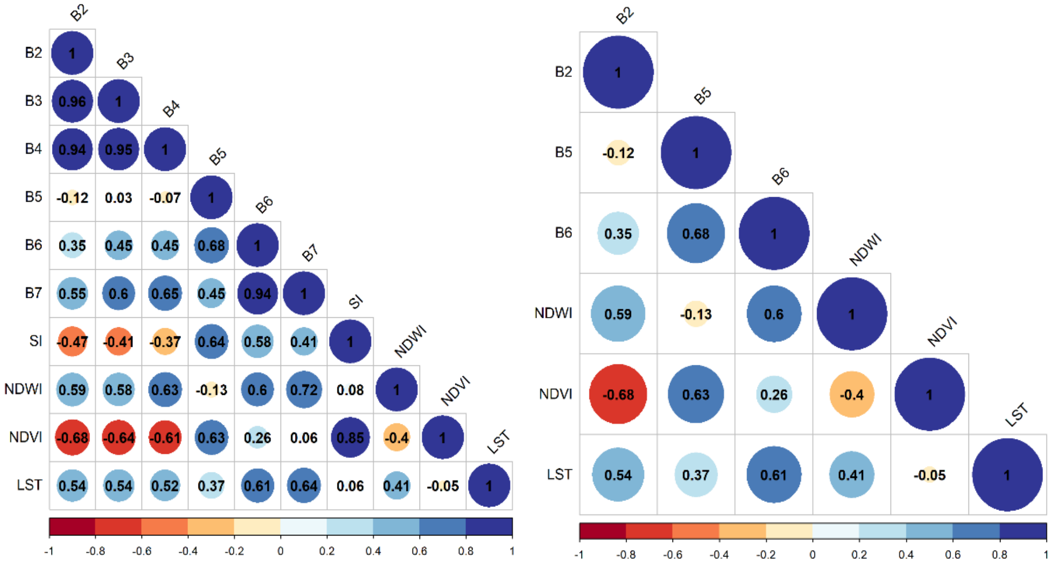

Fig. 4. Results of multicollinearity test to reduce redundancy of SR and ancillary data before importing to a machine learning model.

(B2 = Blue band, B3 = Green band, B4 = Red band, B5 = Near Infrared band, B6 = SWIR-1 band, B7 = SWIR-2 band, NDVI = Normalized Difference Vegetation Index, SI = Normalized Differ-ence Soil Index, NDWI = Normalized Difference Water Index, NDMI = Normalized Difference Moisture Index, LST = Land Surface Temperature)

3.4 Machine Learning Model

3.4.1. CART

The classification and regression tree (CART) is a tree model used in the field of machine learning. CART creates a visualized decision tree to predict a categorical and continuing variable and hence this tree does not create classes of dependent variables. Rather than a classification tree, where an input space of variables is split into subspaces, each of them associates with a class of outputs. Dependent variables are the response values from each observation, given a set of independent variables. As a regression tree does not have preset classes, the output at each stage will be a response value from observations of new dependent variables. A minimization step is applied to create a splitting rule in a tree so that the projected sum variance from two nodes will be deduced.

Proposed by Breiman et al. (Breiman, Classification and Regression Trees, 1984), the classification and regression tree is one of the most adopted methods for handling classification and regression problems. A CART model employs the Gini and least-squared method to deal with categorical and numerical problems, respectively (Breiman, 1996). Given the p^(th) sample is defined as (I_(p,1), I_(p,2),……I_(p,n)…. O_p), where I_(p,n) is a value of the p^(th) sample with "n" features, and "O_p" is an output value of the sample, minimization of the least-squared deviation under Equation (1) will help create a choice to split up a given tree into branches for a CART regression problem.

$$\frac{1}{N}\sum_{V\in U_r}\left(O_p-\bar{O}_r\right)^2+\frac{1}{N}\sum_{V\in U_l}\left(O_p-\bar{O}_l\right)^2 \qquad 6)$$

where "U_r " and "U_l" are training data sets of right and left child nodes, and "N" is the total number of training samples, the outcome of the right and left nodes is denoted as $\bar{O}$ and ($\bar{O}_l$).

3.4.2. Random Forest

A random forest contains a set of tree classifiers, each of them is generated by using a random vector sampled discretely from a given input vector with a vote for the most popular class for categorizing an input vector (Breiman, 2001). The random forest then produces a tree by picking or combining features at each node separately. For each feature or combination of features, a technique called bagging to generate a dataset for training by choosing N instances for replacement randomly, given that N is the size of the original training set, is employed. Any instance is characterized by choosing the class having the highest voting score from all tree predictors within the forest. The design process of decision tree requires a selection measure and a pruning method. Several ways to pick characteristics for decision tree induction are available, and most tactics can clearly measure the attributes. The Information gain ratio and Gini index are mostly adopted attributes as selection metrics for induction of a decision tree. The random forest employs the Gini index as an attribute selection matrix which measures the impurity of an attribute about the classes. As depicted in Equation (7), for a given training set T, the Gini index is expressed as follows:

$$\sum\sum_{j\neq i} (\boldsymbol{f}(\boldsymbol{C_i},\boldsymbol{T})/|\boldsymbol{T}|)\big(\boldsymbol{f}(\boldsymbol{C_j},\boldsymbol{T})/|\boldsymbol{T}|\big) \tag{7}$$

where f(C$_i$,T)/|T| is a probability that a selected instance fits into class C$_i$.

A tree will be formulated each time to its maximum depth by utilizing a mix of features with new training data. The most mature tree remains as-is, this is a benefit of the random forest over other decision tree techniques. Findings reveal that pruning strategies impact the performance of tree-based classifiers rather than the attribute selection criteria (Pal & Mather, 2003).

3.4.3. Support Vector Machine

Introduced by Vapnik et al. (Vapnik, Golowich, & Smola, 1996), support vector machine (SVM) is a supervised classification method to reshape a non-linear environment into a linear one, and make a simple class computable through the generation of a hyper-plane. A kernel function is a mathematical function for transforming data. SVM uses a training dataset to transform an original input into a high-dimensional feature space. A hyper-plane is made from the points of tree classes in the original space of n coordinates. SVM computes the maximum difference across classes to form a classification hyper-plane at the center of the maximum margin. That is, if a point is above the hyper-plane, it is considered as +1; otherwise, it is treated as -1. Support vectors are the training points nearest to the hyper-plane. The new data can then be classified after the decision surface is obtained; such decision surfaces can be utilized to classify auxiliary data. The method is defined over a vector space. The decision surface for a linearly separable space is a hyper-plane, which can be expressed as per Equation (8):

$$wx + b = 0 \tag{8}$$

A vector “w” and constant “b” are derived from a training set of linearly separable items and “x” is an object for characterization. SVM can deal with a problem about linearly restricted quadratic programming such as in Equation (9), and the SVM solution is always globally optimal.

$$\mathbf{min}_{\omega}\frac{\mathbf{1}}{\mathbf{2}} \parallel \boldsymbol{\omega} \parallel^{\mathbf{2}} + \boldsymbol{C}\sum_{\boldsymbol{i}}\boldsymbol{\xi_i} \tag{9}$$

with constraints

$$\boldsymbol{y_i}(\boldsymbol{x_i w} + \boldsymbol{b}) \geqslant \mathbf{1} - \boldsymbol{\xi_i}\ \boldsymbol{\xi_i} \geqslant \mathbf{0}, \forall \boldsymbol{i} \tag{10}$$

By performing non-linear mapping for linearly inseparable objects, the original input data is converted into a higher dimensional space, and the linearly separating hyper-plane can be found in a new space without any additional computational complexity or quadratic programming problems by applying a kernel function (Aizerman, 1964). In other words, to compute similarities across vectors in a high-dimensional space for a linearly inseparable problem, the kernel function is applied to reduce those similarities in the original lower dimensional space.

3.5 Preparation of a Training Signature for the Classification of Oil Palm and LULC using a Machine Learning Model

The spectral bands of Landsat 9 and their derived supplementary bands, as well as the ground sample points (representing each forest type and LULC), were opened in an R statistical program by utilizing raster (Hijmans, 2014) and rgdal (Bivand, 2002) packages. Extraction of training points from the stack images made up of all the spectral and derived ancillary bands was used to produce training signatures. Categorization of oil palm and LULC was finally done using a hierarchy. Six input datasets (ID) created for mapping and classifying oil palm and related LULC in Ranong Province were gradually categorized into a classification hierarchy. Each ID was made up of a variety of spectral data (George, Padalia, & Kushwaha, 2014). Blue, NIR and SWIR2 spectral bands were identified using Landsat 9 in the first stage and results were recorded. Next, an auxiliary variable was added to each ID, and its impact on classification accuracy was assessed using the overall accuracy and Kappa coefficient (**Tab. 3**). In this experiment, three machine learning models were used: Random Forest, Support Vector Machine and CART, and an auxiliary variable was examined in terms of its impact.

Table 3.

Surface Reflectance (SR) and Auxiliary Variable

Auxiliary Variable	Description
ID 1	Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1
ID 2	Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI
ID 3	Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI
ID 4	Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI + LST

4. RESULTS AND DISCUSSION

4.1. Overall Accuracy Results and Kappa Coefficient of the Machine Learning Model

Overall Accuracy simply informs us of the percentage of reference locations that were accurately mapped. The Kappa Coefficient is a statistical test for classifying accurately. Kappa measures how well a categorization fared relative to randomly assigned values. **Table 4** shows overall accuracy results and the Kappa coefficient of Land Use/Land Cover classification in Ranong Province. For ID1 (Surface Reflectance (SR) of Blue, Near Infrared and SWIR-1), it was found that the Random Forest model had OA = 0.9103 and KC = 0.8951. OA and KC would increase with the added variables, as ID4 (Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI + LST) resulted in OA = 0.9341 and KC = 0.9239, which were the highest among 12 models. Most interesting was that the CART model under ID2 (Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI) and ID3 (Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI) had OA = 0.8646, 0.9239 and KC = 0.8427, 0.9103, respectively, and were higher than the Random Forest model under ID2 and ID3. Details of the OA and KC values of all models can be found in **table 4** and **figure 5**. Only the parts with the highest OA and KC values of map ID4 of CART and RF are shown in **figure 6**. In terms of Bare Soils, some models had low producer accuracy, particularly in the ID1 model (Surface Reflectance (SR) of Blue, Near Infrared and SWIR-1). However, after adding more variables, it was found that producer accuracy of such models was higher. Other models had mixed results of producer accuracy and user accuracy in land use/land cover classification. In terms of oil palm, it was found that every SVM and RF in ID 4 is the model that produces the best results. Forest, mangrove, built-up area and water bodies, as well. The classification details using SVM in Oil Palm SVM give the best results over RF, but RF gives OA and KC of all classifications better than SVM.

Table 4.

Overall Producer and User Accuracy of Land Use/Land Cover Classification in Ranong Province.

Model			Bare Soil	Oil Palm	Orchard	Forest	Evergreen Forest	Man grove	Rubber	Built Up Area	Water Bodies	**OA**	**KC**
RF	ID1	UA	1.00	1.00	0.44	1.00	0.94	0.89	1.00	1.00	1.00	**0.91**	**0.90**
		PA	0.33	0.67	0.80	1.00	0.85	1.00	1.00	1.00	1.00		
	ID2	UA	0.50	1.00	0.57	1.00	0.67	1.00	0.88	1.00	1.00	**0.86**	**0.83**
		PA	0.50	1.00	0.80	0.90	0.92	1.00	0.50	1.00	1.00		
	ID3	UA	1.00	0.67	0.50	1.00	0.91	1.00	1.00	1.00	1.00	**0.91**	**0.89**
		PA	0.50	1.00	0.83	1.00	0.95	1.00	0.56	1.00	1.00		
	ID4	UA	0.50	1.00	1.00	1.00	0.81	1.00	0.83	1.00	1.00	**0.93**	**0.92**
		PA	1.00	1.00	0.80	1.00	0.93	0.86	0.91	1.00	1.00		
SVM	ID1	UA	0.00	1.00	0.38	1.00	0.67	1.00	0.75	1.00	1.00	**0.82**	**0.79**
		PA	0.00	1.00	0.60	1.00	0.92	1.00	0.23	1.00	1.00		
	ID2	UA	1.00	1.00	0.44	1.00	0.71	1.00	1.00	1.00	1.00	**0.86**	**0.84**
		PA	0.67	1.00	1.00	1.00	0.92	0.91	0.50	1.00	1.00		
	ID3	UA	1.00	1.00	0.25	1.00	0.89	1.00	0.91	1.00	1.00	**0.89**	**0.87**
		PA	0.50	1.00	1.00	1.00	0.89	1.00	0.63	1.00	1.00		
	ID4	UA	1.00	1.00	0.60	1.00	0.93	0.93	0.72	1.00	1.00	**0.88**	**0.87**
		PA	0.33	0.40	1.00	1.00	0.93	1.00	0.81	1.00	1.00		
CART	ID1	UA	1.00	0.67	0.40	1.00	0.81	0.94	1.00	1.00	1.00	**0.88**	**0.86**
		PA	0.33	0.67	0.57	1.00	0.87	1.00	0.81	1.00	1.00		
	ID2	UA	1.00	0.50	0.56	1.00	0.75	1.00	0.81	1.00	1.00	**0.86**	**0.84**
		PA	0.67	0.33	0.71	1.00	0.88	1.00	0.68	1.00	1.00		
	ID3	UA	1.00	1.00	0.44	1.00	0.89	1.00	1.00	1.00	1.00	**0.92**	**0.91**
		PA	1.00	1.00	1.00	1.00	0.89	1.00	0.72	1.00	1.00		
	ID4	UA	0.67	1.00	0.60	1.00	1.00	1.00	0.82	1.00	1.00	**0.91**	**0.90**
		PA	1.00	0.60	0.86	0.90	0.90	1.00	0.90	1.00	1.00		

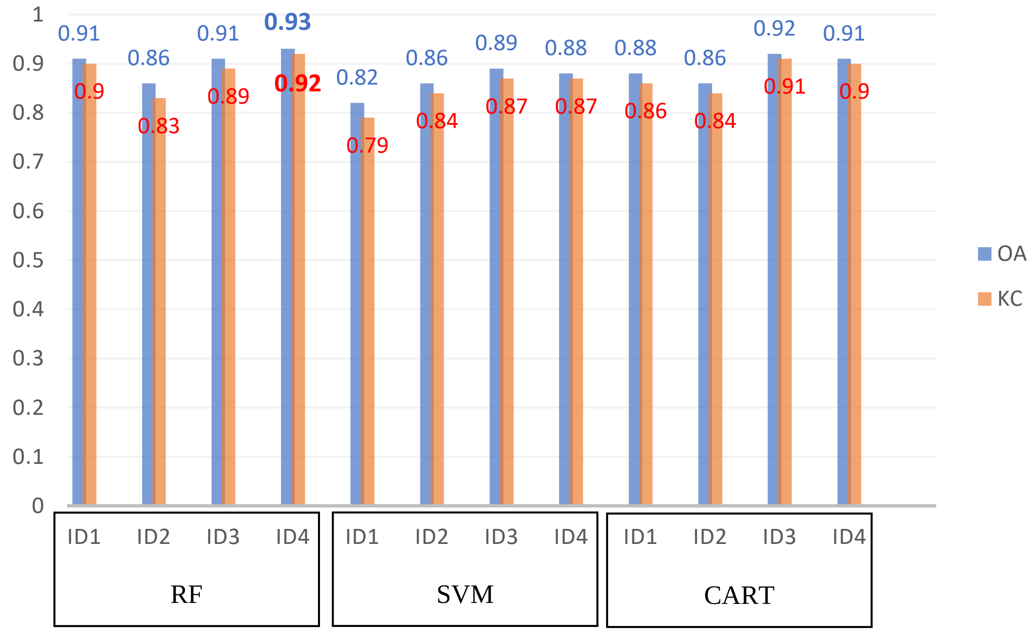

Fig. 5. OA and KC values of all models.

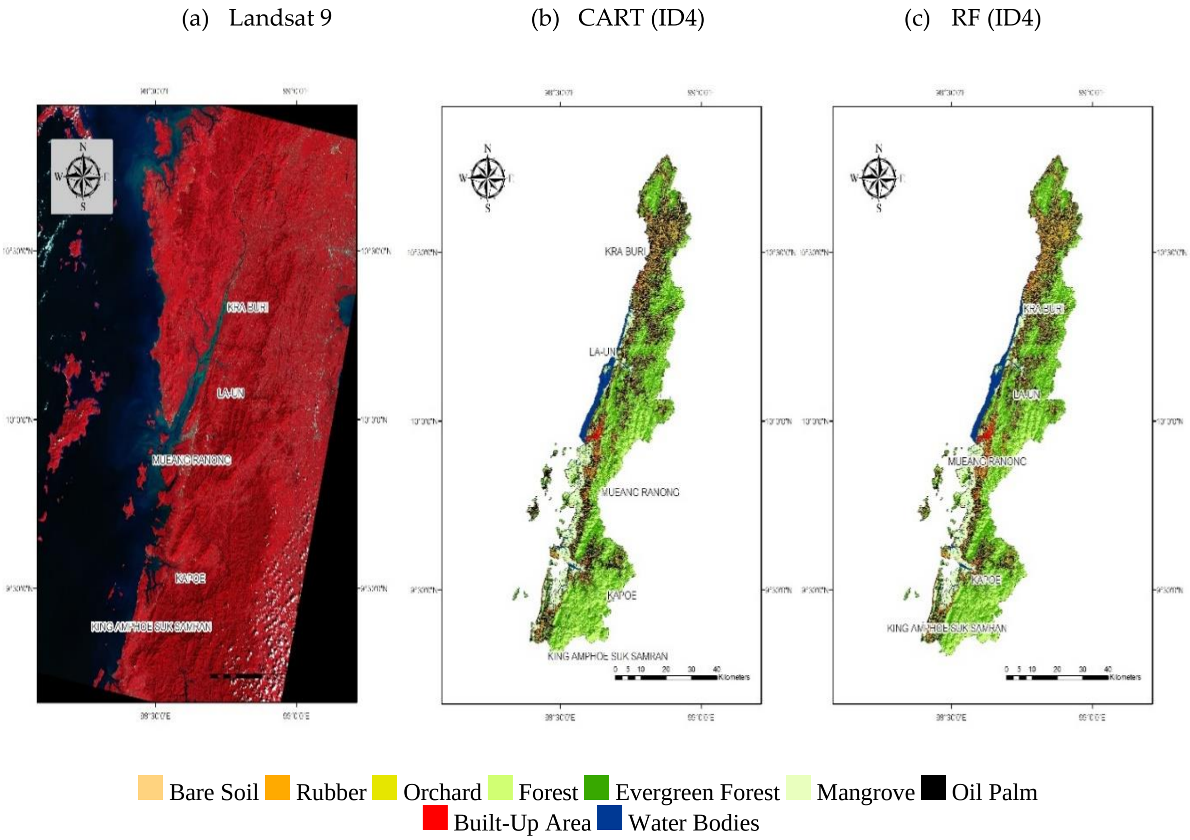

Fig. 6. Example of the improvements of using Surface Reflectance (SR) and Auxiliary Variable (a) Landsat 9 imagery; (b) Results of using CART and Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI + LST; (C) Results of using RF and Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI + LST

Table 5 shows the difference in results of LULC classification on oil palms. Compared to the facts from the Department of Land, ID4 (Surface Reflectance (SR) of Blue, Near Infrared, SWIR-1 + NDWI + NDVI + LST) of CART and RF models gave the classification results closest to the facts from the Department of Land. The reason for choosing these two models was that both OA and KC values were the most accurate. When considered at the district level, it was found that the ID4 of RF models provided classification results of oil palms very close to the facts per the figure of 2.90 sq.km. (1,814 Rai) from the Department of Land, especially for Kra Buri District. The discrepancy between classification of oil palms and actual data was only 83,200 sq.m. (52 Rai). Details of such differences can be found in **table 5**.

Table 5.

Difference in the results of LULC classification on oil palms.

District	Land Development Department	CART (ID4)		RF (ID4)	
		Classified	Difference	Classified	Difference
KRA BURI	29,396	29,174	222	29,448	52
MUEANG RANONG	15,227	15,090	137	12,349	2,878
KAPOE	28,578	30,180	1,602	30,398	1,820
LA-UN	14,438	15,090	652	13,299	1,139
SUK SAMRAN	9,170	11,066	1,896	9,500	330
Ranong Province	96,809	100,599	3,790	94,995	1,814

4.2. Mean Decrease Gini

Mean Decrease Gini (IncNodePurity) - is a measure of variable importance based on the Gini impurity index used for calculating the splits in trees. The first feature and B2 and NDVI shows a high Gini impurity index in the classification model; hence, these features are important for detection of land use/land cover classification, as shown in **figure 7**.

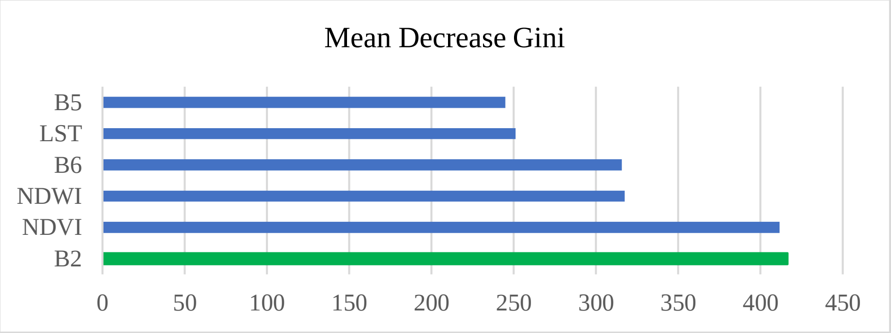

Fig. 7. Mean Decrease Gini of ID4 in a random forest model.

This study demonstrates the capability of using Landsat 9 satellite images for land use/land cover classification. Moreover, using the machine learning model we can precisely classify land use/land cover, especially for oil palms. This study found that multicollinearity is a tool that can significantly reduce variable redundancy. When the remaining variables were used for the Land Use/Land Cover Classification experiment, it was found that variables obtained by fusion data methods such as NDWI, NDVI and LST all resulted in greater accuracy. In addition, machine learning algorithms provide good land use/land cover classification results, especially RF models with the highest OA and KC results. In terms of local economic plants, oil palm, SVM and RF models provided good classification results in both models. This is consistent with Xu's study, which found that Landsat 8 and Sentinel can classify land use/land cover by machine learning (Xu, et al., 2021). Such studies can also classify the life cycle of oil palms using RF as a model, as in this study. In addition, current studies prefer to increase classification accuracy by using the data fusion technique. This technique is a method to bring together different and diversified remote sensing data sources to create new or representative data with the objective of improving data quality, adding more dimensions of data leading to an increase in classification accuracy. This study approach results in more accurate classification than using SR Bands alone. This is consistent with studies by Xu, Shaharum, Poortinga (Xu, et al., 2021; Shaharum N .S., et al., 2020; Poortinga, et al., 2019), especially for Rehman, who found that the multicollinearity test can be used to eliminate factors, removing redundancy of the variables before land use/land cover classification (Rehman, et al., 2021). Besides, Rehman found that adding factors such as indices gradually results in a more accurate model. Like this study, when looking into details, it was found that the SR band 2 NDVI index has more influence on classification, which is consistent with Manandhar's study (Manandhar, Odeh, & Ancev, 2019).

5. CONCLUSIONS

This study highlights the benefits of using more than one data source to create a higher quality dataset and found that machine learning can classify plantations very well. Other researchers can apply such an approach to study other plantations in the future. This study found that, currently, land use/land cover classification cannot use only SR bands alone, so a data fusion technique is necessary to create new or representative data with the aim of improving the quality of information. It found that a machine learning model could also classify land use/land cover precisely. The findings of this study are consistent with several previous studies. Further studies may use a data fusion technique with SAR data to come up with detailed information reflecting objects on the surface of Earth, or use Google Earth Engine, a massive, systematically compiled data fusion source to further expand the project's success. Such techniques could also be applied to other plants in the future.

ACKNOWLEDGEMENTS

We would like to express our appreciation to the Institute of Research and Development, Suan Sunandha Rajabhat University for their support and funding, and the Department of Land and the Treasury Department for permitting us to use their data. Moreover, we would like to give special thanks to Suan Sunandha Rajabhat University, for their support of data collection tools and facilitation throughout this research.

REFERENCES

Adriano, B., Xia, J., Baier, G., Yokoya, N., & Koshimura, S. (2019). Multi-Source Data Fusion Based on Ensemble Learning for Rapid Building Damage Mapping during the 2018 Sulawesi Earthquake and Tsunami in Palu, Indonesia. Remote Sensing, 11(7), 886. doi:10.3390/rs11070886

Aizerman, M. A. (1964). Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. Automation and Remote Control. Automation and Remote Control, 25, 821-837.

Bivand , R. (2002). Spatial econometrics functions in R: Classes and methods. Journal of Geographical Systems volume, 4, 405–421. doi:10.1007/s101090300096

Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123–140. doi:10.1007/BF00058655

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. doi:10.1023/A:1010933404324

Breiman, L. (1984). Classification and Regression Trees. New York: Routledge. doi:doi.org/10.1201/9781315139470

Cheng, Y., Yu, L., Cracknell, A. P., & Gong, P. (2016). Oil palm mapping using Landsat and PALSAR: a case study in Malaysia. International Journal of Remote Sensing, 37(22), 5431-5442. doi:10.1080/01431161.2016.1241448

Cheng, Y., Yu, L., Xu, Y., Lu, H., Cracknell, A. P., Kanniah, K., & Gong, P. (2018). Mapping oil palm extent in Malaysia using ALOS-2 PALSAR-2 data. International Journal of Remote Sensing, 39(2). doi:10.1080/01431161.2017.1387309

Deng, Y., Wu, C., Li, M., & Chen, R. (2015). RNDSI: A ratio normalized difference soil index for remote sensing of urban/suburban environments. International Journal of Applied Earth Observation and Geoinformation, 39, 40-48. doi:10.1016/j.jag.2015.02.010

Du, Z., Li, W., Zhou, D., Tian, L., Ling, F., Wang, H., & Gui, Y. (2014). Analysis of Landsat-8 OLI imagery for land surface water mapping. Remote Sensing Letters, 5(7), 672-681. doi:10.1080/2150704X.2014.960606

George, R., Padalia, H., & Kushwaha, S. P. (2014). Forest tree species discrimination in western Himalaya using EO-1 Hyperion. International Journal of Applied Earth Observation and Geoinformation, 28, 140-149. doi:10.1016/j.jag.2013.11.011

Gutiérrez-Vélez, V. H., & DeFries, R. (2013). Annual multi-resolution detection of land cover conversion to oil palm in the Peruvian Amazon. Remote Sensing of Environment, 129, 154-167. doi:10.1016/j.rse.2012.10.033

Hernández, F. W., Calderón, N. G., & da Silva, P. R. (2022). Oil Palm Yield Estimation Based on Vegetation and Humidity Indices Generated from Satellite Images and Machine Learning Techniques. AgriEngineering(4), 279–291. doi:10.3390/agriengineering4010019

Hijmans, R. J. (2014). Introduction to the ’raster’ package (version 2.2-31).

Huete, A., Justice, C., & Liu, H. (1994). Development of vegetation and soil indices for MODIS-EOS. Remote Sensing of Environment, 49(3), 224-234. doi:10.1016/0034-4257(94)90018-3

Jiménez-Muñoz, J. C., Sobrino, J. A., Skoković, D., Mattar, C., & Cristóbal, J. (2014). Land Surface Temperature Retrieval Methods From Landsat-8 Thermal Infrared Sensor Data. IEEE Geoscience and Remote Sensing Letters, 11(10), 1840-1843. doi:10.1109/LGRS.2014.2312032

Kulpanich, N., Worachairungreung, W., Waiyasusri, K., Saengow, P., & Pinasu, D. (2022). Height Measurement and Oil Palm Yield Prediction Using Unmanned Aerial Vehicle (UAV) Data to Create Canopy Height Model (CHM). Geographia Technica, 17(2), 164-178. doi: 10.21163/GT_2022.172.14

Land Development Department. (2021). Appropriate agricultural promotion guidelines according to the proactive agricultural map database (Agri-Map).

Li, W., Dong, R., Fu, H., & Yu, L. (2019). Large-Scale Oil Palm Tree Detection from High-Resolution Satellite Images Using Two-Stage Convolutional Neural Networks. Remote Sensing, 11, 1-11. doi:10.3390/rs11010011

Liu, M., Hu, S., Ge, Y., & Heuvel, G. B. (2021). Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. Spatial Statistics, 42, 100461. doi:10.1016/j.spasta.2020.100461

Manandhar, R., Odeh, I. O., & Ancev, T. (2019). Improving the Accuracy of Land Use and Land Cover Classification of Landsat Data Using Post-Classification Enhancement. Remote Sensing, 1(3), 330-344. doi:10.3390/rs1030330.

McFEETERS, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7), 1425-1432. doi:10.1080/01431169608948714

Miettinen, J., Shi, C., Tan, W. J., & Liew, S. C. (2012). 2010 land cover map of insular Southeast Asia in 250-m spatial resolution. Remote Sensing Letters, 3(1), 11-20. doi:10.1080/01431161.2010.526971

Mohan, M., Silva, C. A., Klauberg, C., Jat, P., Catts, G., Cardil , A., . . . Dia , M. (2017). Individual Tree Detection from Unmanned Aerial Vehicle (UAV) Derived Canopy Height Model in an Open Canopy Mixed Conifer Forest. Forests, 9(8), 340. doi:10.3390/f8090340

Morel, A. C., Fisher, J. B., & Malhi, Y. (2012). Evaluating the potential to monitor aboveground biomass in forest and oil palm in Sabah, Malaysia, for 2000–2008 with Landsat ETM+ and ALOS-PALSAR. International Journal of Remote Sensing, 33(11), 3614-3639. doi:10.1080/01431161.2011.631949

Mubin, N. A., Nadarajoo, E., Shafri, H. Z., & Hamedianfar, A. (2019). Young and mature oil palm tree detection and counting using convolutional neural network deep learning method. International Journal of Remote Sensing, 40(19), 7500-7515. doi:10.1080/01431161.2019.1569282

Nooni, I. K., Duker, A. A., Van Duren, I., Addae-Wireko, L., & Osei Jnr, E. M. (2014). Support vector machine to map oil palm in a heterogeneous environment. International Journal of Remote Sensing, 35(13), 4778-4794. doi:10.1080/01431161.2014.930201

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86(4), 554-565. doi:10.1016/S0034-4257(03)00132-9

Piekutowska, M., Niedbała, G., Piskier, T., Lenartowicz , T., Pilarski, K., Wojciechowski , T., . . . Kosacka , A. C. (2021). The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. Agronomy, 5(11), 885. doi:10.3390/agronomy11050885

Poortinga, A., Tenneson, K., Shapiro, A., Nquyen, Q., Aung , K. S., Chishtie, F., & Saah, D. (2019). Mapping Plantations in Myanmar by Fusing Landsat-8, Sentinel-2 and Sentinel-1 Data along with Systematic Error Quantification. Remote Sensing, 11(7), 831. doi:10.3390/rs11070831

Rehman, A. U., Ullah, S., Liu, Q., & Khan, M. S. (2021). Comparing different space-borne sensors and methods for the retrieval of land surface temperature. Earth Science Informatics, 14, 985–995. doi:10.1007/s12145-021-00578-6

Santos, V. G., Cuxart , J., Villagrasa, D. M., Jiménez , M. A., & Simó, G. (2018). Comparison of Three Methods for Estimating Land Surface Temperature from Landsat 8-TIRS Sensor Data. Remote Sensing, 10(9), 1450. doi:10.3390/rs10091450

Shafri, H. Z., Hamdan, N., & Saripan, M. I. (2011). Semi-automatic detection and counting of oil palm trees from high spatial resolution airborne imagery. International Journal of Remote Sensing, 32(8), 2095-2115. doi:10.1080/01431161003662928

Shaharum, N. S., Shafri, H. Z., Ghani, W. W., Samsatli, S., Al-Habshi, M. M., & Yusuf, B. (2020). Oil palm mapping over Peninsular Malaysia using Google Earth Engine and machine learning algorithms. Remote Sensing Applications: Society and Environment, 17, 100287. doi:10.1016/j.rsase.2020.100287

Shen, Y., Mercatoris, B., Cao, Z., Kwan, P., Guo, L., Yao, H., & Cheng, Q. (2022). Improving Wheat Yield Prediction Accuracy Using LSTM-RF Framework Based on UAV Thermal Infrared and Multispectral Imagery. Agriculture, 6(12), 892. doi:10.3390/agriculture12060892

Sitthi, A., Nagai, M., Dailey, M., & Ninsawat, S. (2016). Exploring Land Use and Land Cover of Geotagged Social-Sensing Images Using Naive Bayes Classifier. Sustainability, 8(9), 921. doi:10.3390/su8090921

Srestasathiern , P., & Rakwatin, P. (2014). Oil Palm Tree Detection with High Resolution Multi-Spectral Satellite Imagery. Remote Sensing, 6, 9749-9774. doi:10.3390/rs6109749

Sumathi, S., Chai, S. P., & Mohamed, A. R. (2008). Utilization of oil palm as a source of renewable energy in Malaysia. Renewable and Sustainable Energy Reviews, 9(12), 2404-2421. doi:10.1016/j.rser.2007.06.006.

Thenkabail, P. S., Stucky, N., Griscom, B. W., Ashton, M. S., Diels, J., van der Meer, B., & Enclona, E. (2004). Biomass estimations and carbon stock calculations in the oil palm plantations of African derived savannas using IKONOS data. International Journal of Remote Sensing, 25(23), 5447-5472. doi:10.1080/01431160412331291279

Tsouros, D. C., Bibi, S., & Sarigiannidis, P. G. (2019). A Review on UAV-Based Applications for Precision Agriculture. Information (10), 349. doi:10.3390/info10110349

Vapnik, V., Golowich, S. E., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. the 9th International Conference on Neural Information Processing Systems (pp. 281–287). Massachusetts: MIT Press. doi:10.5555/2998981.2999021

Wang, C., Morgan, G., & Hodson, M. E. (2021). sUAS for 3D Tree Surveying: Comparative Experiments on a Closed-Canopy Earthen Dam. Forest, 695. doi:10.3390/f12060659

Worachairungreung, M., Ninsawat, S., Witayangkurn, A., & Dailey, M. N. (2021). Identification of Road Traffic Injury Risk Prone Area Using Environmental Factors by Machine Learning Classification in Nonthaburi, Thailand. Sustainability, 13(7), 3907. doi:10.3390/su13073907

Worachairungreung, M., Thanakunwutthirot, K., & Ninsawat, S. (2021). A Study on Estimating Land Value Distribution for the Talingchan District, Bangkok Using Points-of-Interest Data and Machine Learning Classification. Applied Sciences, 11(22), 11029. doi:10.3390/app112211029

Xu, K., Qian, J., Hu, Z., Duan, Z., Chen, C., Liu, J., . . . Xing, X. (2021). A New Machine Learning Approach in Detecting the Oil Palm Plantations Using Remote Sensing Data. Remote Sensing, 13(2), 236. doi:10.3390/rs13020236

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. International Journal of Remote Sensing, 27(14), 3025-3033. doi:10.1080/01431160600589179

Yin, Q., Liu, M., Cheng, J., Ke, Y., & Chen, X. (2019). Mapping Paddy Rice Planting Area in Northeastern China Using Spatiotemporal Data Fusion and Phenology-Based Method. Remote Sensing, 11(14), 1699. doi:10.3390/rs11141699

Zhou, J., Li, J., Zhang, L., Hu, D., & Zhan, W. (2012). Intercomparison of methods for estimating land surface temperature from a Landsat-5 TM image in an arid region with low water vapour in the atmosphere. International Journal of Remote Sensing, 2582-2602. doi:10.1080/01431161.2011.617396

Zoungrana, B. J.-B., Conrad, C., Amekudzi, L. K., Thiel, M., Da, E. D., Forkuor, G., & Löw, F. (2015). Multi-Temporal Landsat Images and Ancillary Data for Land Use/Cover Change (LULCC) Detection in the Southwest of Burkina Faso, West Africa. Remote Sensing, 7(9), 12076-12102. doi:10.3390/rs70912076.