Manuscript Details

Manuscript number	RSASE_2019_318_R1
Title	Oil Palm Mapping Over Peninsular Malaysia Using Google Earth Engine and Machine Learning Algorithms
Article type	Research Paper

Abstract

Oil palm plays a pivotal role in the ecosystem, environment, economy and without proper monitoring, uncontrolled oil palm activities could contribute to deforestation that can cause high negative impacts on the environment and therefore, proper management and monitoring of the oil palm industry are necessary. Mapping the distribution of oil palm is crucial in order to manage and plan the sustainable operations of oil palm plantations. Remote sensing provides a means to detect and map oil palm from space effectively. Recent advances in cloud computing and big data allow rapid mapping to be performed over large a geographical scale. In this study, 30 m Landsat 8 data were processed using a cloud computing platform of Google Earth Engine (GEE) in order to classify oil palm land cover using non-parametric machine learning algorithms such as Support Vector Machine (SVM), Classification and Regression Tree (CART) and Random Forest (RF) for the first time over Peninsular Malaysia. The hyperparameters were tuned, and the overall accuracy produced by the SVM, CART and RF were 93.16%, 80.08% and 86.50% respectively. Overall, the SVM classified the 7 classes (water, built-up, bare soil, forest, oil palm, other vegetation and paddy) the best. However, RF extracted oil palm information better than the SVM. The algorithms were compared and the McNemar's test showed significant values for comparisons between SVM and CART and RF and CART. On the other hand, the performance of SVM and RF are considered equally effective. Despite the challenges in implementing machine learning optimisation using GEE over a large area, this paper shows the efficiency of GEE as a cloud-based free platform to perform bioresource distributions mapping such as oil palm over a large area in Peninsular Malaysia.

Keywords	cloud computing; Landsat; oil palm

Corresponding Author	Helmi Shafri
Corresponding Author's Institution	Universiti Putra Malaysia (UPM)
Order of Authors	Nur Shafira Nisa Shaharum, Helmi Shafri, Wan Azlina Wan Ab Karim Ghani, Sheila Samsatli, Mohammed Al-Habshi, Badronnisa Yusuf

Oil Palm Mapping Over Peninsular Malaysia Using Google Earth Engine and Machine Learning Algorithms

Nur Shafira Nisa Shaharum[a], Helmi Zulhaidi Mohd Shafri[a,b*], Wan Azlina Wan Ab Karim Ghani[c], Sheila Samsatli[d], Mohammed Mustafa Abdulrahman Al-Habshi[a] and Badronnisa Yusuf[a]

[a]Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia; [b]Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia; [c]Department of Chemical and Environmental Engineering/Sustainable Process Engineering Research Centre (SPERC), Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia; [d]Department of Chemical Engineering, University of Bath, Claverton Down, BA2 7AY, United Kingdom.

Corresponding author: helmi@upm.edu.my

Oil Palm Mapping Over Peninsular Malaysia using Google Earth Engine and Machine Learning Algorithms

Abstract

Oil palm plays a pivotal role in the ecosystem, environment, economy and without proper monitoring, uncontrolled oil palm activities could contribute to deforestation that can cause high negative impacts on the environment and therefore, proper management and monitoring of the oil palm industry are necessary. Mapping the distribution of oil palm is crucial in order to manage and plan the sustainable operations of oil palm plantations. Remote sensing provides a means to detect and map oil palm from space effectively. Recent advances in cloud computing and big data allow rapid mapping to be performed over large a geographical scale. In this study, 30 m Landsat 8 data were processed using a cloud computing platform of Google Earth Engine (GEE) in order to classify oil palm land cover using non-parametric machine learning algorithms such as Support Vector Machine (SVM), Classification and Regression Tree (CART) and Random Forest (RF) for the first time over Peninsular Malaysia. The hyperparameters were tuned, and the overall accuracy produced by the SVM, CART and RF were 93.16%, 80.08% and 86.50% respectively. Overall, the SVM classified the 7 classes (water, built-up, bare soil, forest, oil palm, other vegetation and paddy) the best. However, RF extracted oil palm information better than the SVM. The algorithms were compared and the McNemar's test showed significant values for comparisons between SVM and CART and RF and CART. On the other hand, the performance of SVM and RF are considered equally effective. Despite the challenges in implementing machine learning optimisation using GEE over a large area, this paper shows the efficiency of GEE as a cloud-based free platform to perform bioresource distributions mapping such as oil palm over a large area in Peninsular Malaysia.

Keywords: cloud computing; image classification; Landsat; machine learning; oil palm

1. Introduction

Malaysia is a Southeast Asian country sharing borders with Thailand, Indonesia and Brunei. Malaysia is a tropical country with two geographical regions: Peninsular Malaysia and Borneo (Sabah and Sarawak). It experiences a humid, hot and rainy climate throughout the year, experiencing temperature ranging between 23°C – 32°C throughout the country. This allows Malaysia to generate income from agricultural crop activities such as paddy cultivation, rubber and oil palm planting (Fahmi et al. 2013; Nambiappan et al. 2018).Among the agricultural crops, oil palm produces the highest amount of biomass, and as one of the largest palm oil exporters in the world, the total number oil palms planted in Malaysia reached over 5 million hectares (ha) in 2017 (Ng et al. 2012). Moreover, oil palm was the main contributor of agricultural crops to the country's GDP in 2017 with a total contribution of 46.6% (Mahidin 2018). Despite its benefits, oil palm activities contributed to massive deforestation and caused negative impacts to the environment (Fitzherbert et al. 2008) and therefore, oil palm activities have been labelled as the main threat to the earth by contributing to the global warming and climate change (Shuit et al. 2009). Destroying wildlife habitats and forests for planting oil palm trees have worsened the negative implications. Even though palm oil can be used as a renewable energy source and help contribute to the 17 Sustainable Development Goals as presented by the United Nations, it is important to note that the environment will be in jeopardy without proper management and monitoring on the oil palm industry, which will in turn affect environmental sustainability. However, managing huge areas of oil palm plantations will be challenging. Furthermore, implementing ground surveys or other traditional survey methods will require a tremendous amount of time, effort and high cost. A number of people are required to execute data collection over a large area and therefore, high computational power will be essential to process such big data. Hence, the utilisation of remote sensing is a suitable and a cost-effective method for collecting data covering a huge area.

The use of remote sensing for oil palm applications can be found in many publications using a variety of sensors, platforms and algorithms. For example, Thenkabail et al. (2004) used four bands with 4 m of spatial resolution from IKONOS to carry out a study on oil palm biomass estimations and carbon stock calculations. Before implementing image classification, the band was first masked by extracting the oil palm feature from non-oil palms. Next, Gutiérrez-Vélez and DeFries (2013) utilised MODIS data with 250 m of pixel size and successfully produced an oil palm map covering an area of 939,204 km^2. Another similar study using MODIS data was conducted on a larger study area covering several regions in Southeast Asia including Peninsular Malaysia, Sumatra, Java, Borneo, Sulawesi and Mindanao. The study has successfully classified a total of 13 classes together with mangrove forests, rainforests and large-scale palm plantations (Miettinen et al. 2012). This indicated that studies using coarse spatial resolution can be implemented in oil palm studies. On the other hand, using higher spatial resolution data, analysis on oil palm studies can be improved and more information can be extracted. Jusoff and Pathan (2009) and Shafri and Hamdan (2009) used hyperspectral sensor to map individual oil palm trees. A more subtle analysis was conducted by Shafri et al. (2011) via Maximum Likelihood Classifier (MLC) and successfully detected Ganoderma disease infections in the plantations with an overall accuracy of 82%. More recently, the high spectral resolution as provided by hyperspectral data has allowed Camacho et al. (2019) to successfully produce an oil palm map distinguishing healthy from diseased palm trees.

In terms of classification algorithms, Morel et al. (2012) have successfully distinguished between forest and oil palm areas on Landsat data using k-means and MLC algorithms. Then, Glinskis and Gutiérrez-Vélez (2019) used MLC algorithm to classify oil palm and successfully categorised it into 3 stages (infant palm, juvenile palm and adult palm) via Sentinel 1 and 2 data. Studies using more advanced algorithms or approaches such as Support Vector Machine (SVM), Random Forest (RF), Deep Learning, Artificial Neural Network (ANN) and other machine learning algorithms to classify oil palm land cover tend to produce better results (Nooni et al. 2014; Li et al. 2015; Lee et al. 2016; Noi and Kappas 2018). For example, Cheng et al. (2016) performed land cover classifications on Landsat and

ALOS-PALSAR remote sensing data via SVM and Minimum Distance algorithms. The classifications were applied on two different sites, and the overall accuracies produced by SVM were higher than Minimum Distance for both Landsat and ALOS-PALSAR data. Cheng et al. (2018) expanded the oil palm classification on larger areas covering Malaysia, Indonesia, Thailand, Nigeria and Ghana using ALOS-PALSAR data. The study achieved an overall accuracy of more than 94% for all the aforementioned countries. A review of studies on fusion techniques between optical and radar data to map land use was conducted by Joshi et al (2016) and it showed that fusion techniques are efficient for cloud issue. De Alban et al (2018) combined Landsat and L-band Synthetic Aperture Radar (SAR) data to carry out land use land cover change application in tropical landscapes. A machine learning RF algorithm was used to classify the land use and the accuracy obtained was 92.96% to 93.83%. However, fusion technique requires large amount of time and data to produce the cloud-free image.

As shown above, there have been several oil palm studies using various remote sensing data, however, most of the studies were limited to small areas (Li et al. 2016; Chong et al. 2017; Charters et al. 2019; Fawcett et al. 2019) and utilised personal computers, requiring the ability to store data and perform image processing using remote sensing software that were mostly commercial. Data obtained from the impacts of oil palm activities that were conducted on small areas are not suitable and insufficient to be used for measuring the sustainability level for a huge area, especially for the whole country. On the other hand, the utilisation of very high-resolution images on big areas will be costly in addition to requiring high computational power which will be essential to process the data. Even so, GEE cloud computing provides an alternative to process huge amount of geospatial data with zero cost and without the need to personally store the data on the personal computer. Sidhu et al. (2018) performed land cover change analysis in Singapore's landmass via GEE and the result showed that the forest cover was affected by the monsoon cycles. Another study by Oliphant et al. (2019) to map cropland over Southeast and Northeast Asia via Landsat was carried out using GEE, but no specific crops (e.g. oil palm, paddy and others) are mapped. In Malaysia, first ever effort to map oil palm over the Peninsular Malaysia using cloud computing platform was done using the Remote Ecosystem Monitoring Assessment Pipeline (REMAP) tool as conducted by Shaharum et al. (2019). However, it was limited to only the use of RF classifier, limiting the investigation of the performance of other machine learning algorithms. Furthermore, as the toolbox is not programmable, the parameters of the classifier cannot be optimized or tuned accordingly. In addition, the imagery data used in REMAP was fixed and cannot be filtered to produce the best cloud-free data. In addition, the GEE algorithms were run in code editor module, allowing more optimization parameters to be tested for classification. To the best of our knowledge, there has been no report on the utilisation of GEE for oil palm mapping over the entire Peninsular Malaysia. Hence, this study was conducted to test the capability of 30 m Landsat data using the GEE cloud computing platform and compare machine learning algorithms such as SVM, CART and RF to map oil palm land cover over Peninsular Malaysia covering an area of 132,265 km^2. Even though there are many available remote sensing data and techniques available to classify the oil palm plantation, selecting the best technique will be vital.

2. Material and Methods

2.1 Study area

Malaysia is located between Thailand, Singapore and Indonesia. It comprises of two regions: Peninsular Malaysia and Borneo (Sabah and Sarawak). This study covers Peninsular Malaysia (N 4°00'0.00", E 102°29'59.99"; Fig 1) or West Malaysia with an approximate land area of 132,265 km^2. Malaysia is a tropical country experiencing both hot and humid weather throughout the year. The temperature ranges between 23°C – 32°C and during hot weather, the temperature can exceed over 40°C (Shahar 2016).

[Figure 1 near here]

2.2 Google Earth Engine

Vast amount of the geospatial remote sensing data provided in the GEE has allowed the powerful cloud-based platform to be used in various studies involving deforestation, oil palm plantations, environmental assessment, change detection and urban classifications (Patel et al. 2015; Dong et al. 2016; Goldblatt et al. 2016; Shelestov et al. 2017). GEE can be accessed either through Application Programming Interface (API) or web-based Interactive Development Environment (IDE) (Gorelick et al. 2017). The data catalog provided in the GEE houses a multi-petabyte accessible geospatial dataset that is made up of Earth-observing remote sensing images, including Landsat, MODIS, Sentinel-1 and Sentinel-2.

[Figure 2 near here]

Figure 2 shows the GEE platform via Javascript API and it allows the user to control the data through coding. The user can write the programs using client libraries in Python and Javascript (programming languages). Furthermore, the client libraries provide objects for Images, Collections and other data types. In fact, the user can perform various remote sensing analyses in the GEE API platform such as image classifications, multitemporal urban extents, post-processing and object detection. Enormous amount of Earth Engine public data catalog provided in the cloud-based GEE platform helps the user to process very large geospatial datasets without having to suffer the information technology pains including the need of high computational power resources and huge amount of storage.

2.3 Data collection and pre-processing

The availability of 30 m Landsat 8 images for the study area were obtained from the United States Geological Survey through the GEE platform (Roy et al. 2014). The images were already being pre-processed and corrected at Top-Of-Atmosphere (TOA) reflectance as explained by Chander et al. (2009) by converting at sensor (spectral radiance) to exoatmospheric TOA reflectance. The benefits of using images that have been corrected at TOA reflectance are it compensates for different values of the exoatmospheric solar irradiance occur from spectral band differences and the TOA reflectance can eliminate the cosine effect of different solar zenith angles due to the time difference between data acquisitions. Also, it corrects the dissimilarity in the Earth–Sun distance between different data acquisition dates.

[Table 1 near here]

This study utilised only 7 bands of Landsat 8 with 30 m spatial resolution (see Table 1). Landsat 8 data is obtained via passive remote sensing, and it is sensitive towards clouds. Several Landsat 8 images taken from year 2016 and 2017 were patched together to attain the missing information that were blocked by the clouds. The existence of clouds can affect the quality of the remote sensing data and furthermore, the information beneath the cloud will be unclassified. The utilisation of commercial remote sensing software to perform image patching on a huge area consumes significant resources and time (Gambo et al. 2018;

Shaharum et al. 2018). However, the GEE platform allows the user to perform data acquisition and image patching in a few seconds. Furthermore, it allows the user to set the percentage of the cloud cover and the desired date of the satellite data to be used.

3. Methodology

Several geospatial datasets were utilised in this study to produce the oil palm land cover maps over Peninsular Malaysia: (i) 30 m Landsat 8 data from 2016 to 2017 (7 original bands) (ii) Shuttle Radar Topographic Mission (SRTM), Digital Elevation Model (DEM), (iii) Additional data including NDVI, Normalised Difference Water Index (NDWI) and others. The workflow adapted for this study is shown in Figure 3.

[Figure 3 near here]

This study compared 3 different machine learning algorithms (SVM, RF and CART), and a total of 7 classes including oil palm were classified. The importance of oil palm plantation has been discussed in the introduction and therefore, this study focuses on producing oil palm land cover map. Moreover, the land cover map produced can later be used in the next study to assess the impacts of oil palm plantation over Peninsular Malaysia.

3.1 Data used for classification

As illustrated in Table 1, a total of 7 bands obtained from Landsat 8 images were used and these bands were used to generate additional data (see Table 2). A number of equations were used to produce additional data that will be included together with the other 7 bands to be used in the image classification stage.

[Table 2 near here]

The layers in Table 2 were stacked together with Landsat 8 bands (Table 1) to be used in the classification process. These additional layers are capable of extracting a certain information in a more efficient way. For example, NDVI is derived from the ratio between Red and Near-infrared (NIR) reflectance bands. Furthermore, NDVI is sensitive towards chlorophyll content and the green leaf density. The presence of chlorophyll in green vegetations absorbs in the red band. Hence, NDVI is useful to extract information of the green vegetations on the ground (Bro-Jørgensen et al. 2008).

3.2 Sampling

Samples were created in the GEE platform and a total of 7 classes were identified: water, built-up, bare soil, oil palm, forest, other vegetation and paddy. The samples were created using the point format for every state in Peninsular Malaysia, covering a total of 11 states via random sampling. The samples were created with the aid of land cover map provided by the Department of Agriculture (DOA) and high-resolution Google Earth images as shown in Figure 4(a). The samples were then divided into two components: training and testing. A total of 70% from the whole created samples (4307 points) were used to classify the Landsat images and the remaining 30% of the samples (1846 points) were used to validate and assess the accuracy of the algorithms used. The classification and validation were done in GEE and the accuracy assessment was calculated using the common confusion matrix method.

[Figure 4 near here]

3.3 Supervised machine learning algorithms

3.3.1 Support Vector Machine

Supervised classification can be conducted using machine learning and non-machine learning algorithms. SVM is a type of supervised machine learning algorithm that works well in classification and regression. It uses a hyperplane (see Figure 5) to divide the support vectors to distinctly classify the data points, and there are many possible ways for the hyperplane to separate the support vectors in which, the main objective of SVM is to find the hyperplane that has the maximum margin (separate support vectors of both classes at a maximum distance) (Maxwell et al. 2018).

[Figure 5 near here]

SVM comprises of several hyperparameters: kernel type, gamma and penalty value. These hyperparameters can be tuned and adjusted to improve the performance of SVM in image classification.

3.3.2 Classification and Regression Tree

The CART is similar to DT. CART, which is a type of supervised machine learning algorithm that forms a binary decision tree. It involves the identification and construction of the tree using the training samples for which the correct classification is unknown. The decision tree starts with a root node derived from any variable in the feature space and minimises a measure of the impurity of the two sibling nodes (see Figure 6). Then, the decision tree grows by means of the successive subdivisions until it reaches a stage where there is no significant reduction in the measure of impurity when further division is implemented (Bittencourt and Clarke 2003; Jiang et al. 2010).

[Figure 6 near here]

The decision tree is made of multilevel and multi-leaf nodes and the decision tree will undergo a pruning process once it is constructed. The constructed trees are often over-fit because an excessive number of nodes and branches are often being created. Therefore, the tree can be pruned by controlling the parameters or thresholds for the new branches (Calbury 2016).

3.3.3 Random Forest

RF or Random Decision Forest is a non-parametric machine learning algorithm that can be used in both classification and regression analysis. It is a type of ensemble learning algorithm that ensembles a number of decision trees and forms a forest (see Figure 7). This algorithm combines random features or a combination of features at each node to grow a tree. The bagging method is used in this algorithm to generate the training samples, and each selected feature is drawn randomly by the replacement of N (size of original training set) examples. The examples are classified based on the highest voted class produced from all the trees in the forest (Pal 2005).

[Figure 7 near here]

One of the most frequently used attributes in the decision tree induction is the Gini Index. For a given training set T, selecting one case (pixel) at random and assuming that it belongs to some class C_i, the expression can be written as:

$$\sum \sum_{j \neq i} (f(C_i,T)/|T|)((f(C_j,T)/|T|) \qquad (1)$$

where $f(C_i, T)/|T|$ is the probability that the selected case belongs to class C_i. Gini Index acts as an attribute selection measure in RF that measures the impurity of an attribute with respect to the classes. Since RF works by assembling a number of trees, whereby N is to form a forest, the value of N can be defined by the user to get the best output of the classification. The RF algorithm can use a large number of trees in the ensemble and as a result, it works well in high dimensional data (Gislason et al. 2006).

3.3.4 Hyperparameters optimisation

Every algorithm has its own built-in hyperparameters/parameters that can be adjusted and further tuned to improve its performance. A hyperparameter is a parameter which contains a value that is set or defined before performing any learning process, and different model training algorithms consist of different hyperparameters. In this study, the hyperparameters in SVM, CART and RF algorithms were optimised in GEE, and the involved hyperparameters were tabulated in Table 3.

[Table 3 near here]

These hyperparameters are tuneable and can directly affect the robustness of the learning models, thus optimisation of the hyperparameters is required to achieve the best performance level of the algorithms. The identification of the hyperplane in SVM can be due to the type of kernel, k: Linear, Radial basis function, Polynomial and Sigmoid. Kernels are used to solve a non-linear problem in a higher dimension and is usually referred to as kernel trick (Afonja 2017). Gamma, *g* is the hyperparameter in SVM that defines how far the influence of a single training example reaches. A high value of gamma considers only nearby points (near identified hyperplane) in the calculation. Conversely, low gamma considers far away points to be included in the calculation for the separation of the hyperplane (Patel 2017). As for the penalty or regularisation hyperparameter, *C* in SVM is to avoid misclassification in the learning model. A larger value of *C* tells SVM to produce a smaller-margin hyperplane and on the contrary, a small value of *C* enlarges the margin of the hyperplane.

[Figure 8 near here]

In this study, a total of four hyperparameters were fine-tuned in CART. Firstly, the cross-validation factor, *cv* in CART partitions of the training samples were tuned into K (number of folds) equally sized subsamples. Assuming that the training samples were divided into 10 folds of subsamples, 9 of the subsamples are used as training data and the other 1 subsample as validation as shown in Figure 8 (Ivanovic 2016). Then, max depth, *d* is used to determine the maximum of the tree depth in the model. The number of terminal nodes increases proportionally to the depth of the tree. For *d* equals to 1 will have 2 terminal nodes, and *d* equals to 2 will have a maximum of 4 nodes. The maximum of the nodes in a tree depends on the depth of the tree by implementing the rule of 2 to the power of *d* (Molnar 2016). Minimum leaf population is used to set the minimum number of samples for a terminal node (leaf) and minimum split population is set to define the minimum the number of sub-nodes to be divided by a node (Brid 2018). Finally, the average produced using the arithmetic mean and the class probabilities taken from the hyperparameter in RF, namely number of trees, *t* set in the model will be the final classification decision (Pal 2005; Belgiu and Drăguț 2016).

3.4 McNemar's test

McNemar's test is a statistical test that applies to 2 x 2 contingency table. Sometimes, it is known as McNemar's Chi-Square test because it has a chi-square distribution. McNemar's test is conducted to determine whether if there are differences on a dichotomous dependent variable between two related classifiers or groups (Pal et al. 2013). The McNemar's test has been used by Yu et al. (2017) to determine the difference between classifications based on other pairs of features. Duro et al. (2012) performed McNemar's test to compare the classification results between DT, RF and SVM via object-based and pixel-based techniques. The result showed that the p-value via object-based was statistically significant ($p < 0.05$) when comparing DT with either RF or SVM. On the other hand, no statistically significant difference ($p > 0.05$) was produced when comparing the results obtained from different algorithms via pixel-based technique.

4. Results and Discussion

4.1 Land cover classifications

This study was aimed to produce an oil palm land cover map over Peninsular Malaysia by comparing SVM, CART and RF machine learning algorithms in the GEE platform. A total of 7 classes (water, built-up, bare soil, forest, oil palm, other vegetation and paddy) were classified. However, the classification output analysis emphasized only on oil palm because the produced oil palm map will later be used to evaluate the spatial distribution of oil palm in Peninsular Malaysia and will be included into a Geographic Information System (GIS) database for further analysis. The hyperparameters were optimised and classified maps produced by the algorithms are shown in Figure 9.

[Figure 9 near here]

The hyperparameters optimisation was carried out in the GEE. A grid search method was implemented for each algorithm to find the best hyperparameters to be used for the classification (Gupta et al. 2018). Generally, the hyperparameters used to produce the outputs for each algorithm are as shown in Table 3.

In this study, 7 classes were identified in which, water classified features with water elements such as lakes, sea, rivers and ponds. Built-up classified buildings, metals, concretes and roads. Then, bare soil classified features that are bare land, open areas and places full of sand or soil (such as construction site). Oil palm classified oil palm trees while other vegetation classified features other than oil palm and forests such as shrubs, other crops and plantations. Since the aim of this study was to test the performance of machine learning algorithms to extract oil palm plantation from 30 m Landsat 8 images in the cloud-based, GEE platform, additional information such as NDVI, NDWI and slope were included to enhance the classification, especially in distinguishing one class from another. The produced oil palm map provides the information on the oil palm distribution for 2017 and furthermore, the map can later be used in the future studies such as to evaluate the impact of oil palm land cover in detailed.

4.2 Overall accuracies and land cover maps comparison

A total of 30% of testing samples (water: 213, built-up: 311, bare soil: 126, forest: 470, oil palm: 331, other vegetation: 276 and paddy: 119) were used to validate the classified land cover maps, and the overall accuracies obtained for each state were calculated (see Table 4). The total area of oil palm in Peninsular Malaysia produced by CART, RF and SVM were

3005758 ha, 2795287 ha and 2924434 ha respectively. Table 4 indicates that SVM produced the highest overall accuracy with an average of 93.16%. That is followed by the overall accuracies produced by RF and CART with an average of 86.50% and 80.08% respectively. The overall accuracies produced were calculated via confusion matrix based on the accuracies of the 7 classified classes.

[Table 4 near here]

[Table 5 near here]

By referring to the inventory provided by the Malaysia Palm Oil Board (MPOB), the area of oil palm plantations produced by SVM, CART and RF were compared for each state in Peninsular Malaysia. Table 5 shows the difference of oil palm area produced by SVM, CART and RF by comparing the generated results with the MPOB inventory. However, the limitation of 30 m coarse resolution data, RF, CART and SVM have overestimated the oil palm area and misclassified other classes as oil palm land cover. Based on the classified oil palm areas tabulated in Table 5, most of the states overestimated the oil palm areas. Kedah overestimated more than 60000 ha and followed by Selangor with an overestimation of more than 56000 ha of oil palm area. Then, the result showed that at least 1000 ha of land area was misclassified as oil palm in Perlis. This is because the misclassified pixels were due to the similarity of the reflectance value of the pixels. Therefore, the pixels that were misclassified as oil palms have contributed to the overestimation of the oil palm area for the aforementioned states. Conversely, all three algorithms underestimated the oil palm area for Melaka. Overall, all the machine learning algorithms, SVM, RF and CART overestimated the oil palm area for Peninsular Malaysia. Although SVM produced the highest overall accuracy, RF produced the least errors in comparison with the MPOB inventory in oil palm classification by producing an overall error of 0.03%, followed by SVM and CART with 0.08% and 0.11% for the whole oil palm area of Peninsular Malaysia respectively. Furthermore, RF classified oil palm land cover and produced the nearest result (oil palm area) to the MPOB inventory for most of the states: Negeri Sembilan, Pulau Pinang, Kedah, Pahang, Perak, Perlis and Terengganu. Then, CART extracted the most accurate oil palm areas for Johor and Kelantan, and SVM extracted the best for Melaka.

This study had tested the performance of three algorithms (CART, RF and SVM) with fine-tuned hyperparameters on 30 m Landsat data, and managed to produce oil palm land cover maps over Peninsular Malaysia using a cloud-based platform, GEE. The powerful cloud computing platform, GEE has made mapping oil palm land cover over Peninsular Malaysia using Landsat data possible. However, this study has confronted a few setbacks. Firstly, the utilisation of 30 m spatial resolution data might produce errors due to mixed pixels and furthermore, there might be more than one class in a single pixel. Then, the similarity of the reflectance between other vegetation and oil palm as well as between bare soil and built-up had caused confusion in the classification. Furthermore, the images used were the result from image patching, in which the product might contain errors in the pixels, hence reducing the quality of the image. Peninsular Malaysia is a huge area, and to obtain a single cloud-free image for a tropical region covering such huge area is merely impossible. Thus, image patching is one of the alternatives to obtain an almost cloud-free data. Therefore, it is challenging to ensure the quality of optical data, especially when it involves huge tropical area.

[Figure 10 near here]

[Figure 11 near here]

Figures 10 and 11 are the selected focused regions in Pulau Pinang and Selangor respectively that were classified by CART, RF and SVM. The feature (in the red box) showed in Figure 10(a) is oil palm trees. Based on the classified images (Figures 10(b), 10(c) and 10(d)), the result showed that SVM misclassified most of the oil palm pixels as other vegetation. Furthermore, SVM misinterpreted the bare soil pixels with built-up in Selangor as shown in Figure 11(d). On the other hand, the oil palm and bare soil pixels for both areas (Selangor and Pulau Pinang) were found to be well classified by CART and RF. These findings showed that the utilisation of additional layers (NDVI, NDWI and others) in the tree methods implemented in RF and CART is more efficient. In addition, both tree-based algorithms (RF and CART) can classify the pixels better than the SVM that works via maximizing the hyperplane. Moreover, the failure of SVM algorithm in separating the support vectors had led to classification errors. As for RF and CART, the RF algorithm has improved the classification as the trees in the RF were ensembled into a forest, and finally the classes were defined based on the majority vote. Although Figures 10 and 11 showed that SVM misclassified oil palm and bare soil pixels, the best algorithm to classify all the 7 classes for the whole Peninsular Malaysia is SVM. However, by comparing all three machine learning algorithms, this study agreed that RF extracted oil palm class the best for the whole Peninsular Malaysia.

4.3 McNemar's test

McNemar's test has been carried out in this study to measure the significance between the classification of SVM and CART, SVM and RF and CART and RF. The 2 x 2 contingency table as tabulated in Table 6 was used to calculate the p-values.

[Table 6 near here]

The null hypothesis of this test states that the probability of Test 1 being correctly classified is equal to the probability of Test 2 being correctly classified. Also, the probability of Test 1 being incorrectly classified is equal to the probability of Test 2 being incorrectly classified. In other words, Pa + Pb = Pa + Pc or Pb + Pd = Pc + Pd, which leads to Pb = Pc.

Pa = Probability of Test 1 being positive and Test 2 being positive
Pb = Probability of Test 1 being positive and Test 2 being negative
Pc = Probability of Test 1 being negative and Test 2 being positive
Pd = Probability of Test 1 being negative and Test 2 being negative

The p-value will be calculated and the value of $p < 0.05$ is considered as a significant result, thus rejecting the null hypothesis. In this study, calculations of the p-value using the formula demonstrated by Foody (2004) were conducted for all the algorithms and the results are tabulated in Table 8.

[Table 7 near here]

The p-value obtained when comparing between SVM and RF is 0.28 ($p > 0.05$), while the other two comparisons obtained values of $p < 0.05$. Due to the robustness and powerful machine learning algorithms, SVM and RF algorithms can classify the pixels well. Hence, the comparison between SVM and RF gave a non-significant p-value > 0.05 and thus, accepted the null hypothesis.

5. Conclusion

In this study, we utilised 30 m Landsat data in the GEE platform to produce oil palm land cover maps over Peninsular Malaysia. The GEE platform is controllable and it provides options especially in selecting the processing methods, algorithms and data input. Furthermore, it allows users to design the workflow based on their needs. In this study, three machine learning algorithms were used and the hyperparameters were tuned. Accuracy assessments for the classified maps were conducted using high-resolution Google Earth images and the map provided by the DOA. The comparison of the classified oil palm areas with the inventory provided by MPOB has shown that there is a large uncertainty of oil palm land cover in Perlis, Kedah and Selangor. Overall, CART, SVM and RF were able to classify the land cover maps and produced acceptable results by producing an overall accuracy of 80.08%, 93.16% and 86.50% respectively. Then, McNemar's test was conducted and it showed that significant p-values were obtained when comparing CART to both SVM and RF. However, the test showed a non-significant value when comparing between RF and SVM. This shows that both methods can reliably be used to produce high accuracy maps in GEE and later be used to classify other crops. Moving on, such timely and high accuracy estimates of oil palm areas could be embedded with other ancillary GIS data for a variety of monitoring and decision-making applications, including yield prediction, supply-chain logistics, commodity markets, bioenergy estimation and more.

GEE provides various geospatial data including Sentinel 2, Sentinel 1 and MODIS. The utilisation of higher spatial resolution data such as Sentinel 2 with 20 m to 10 m of pixel size can be tested to improve the classification. Moreover, Sentinel 1 works with active sensors, and it is suitable to be used on tropical regions. The integration of Sentinel 1 data in the GEE platform can reduce the time needed to process huge amounts of radar data. On top of that, there are many more methods available in GEE to pre-process remote sensing data, in which some methods might produce good results and able to improve the accuracy of the data. In addition, the programmable platform produces the possibilities for the cloud computing GEE to be integrated with the powerful deep learning methods.

Acknowledgements

We would like to thank Universiti Putra Malaysia for their facilities and support for this research. Sponsorship from the Engineering and Physical Sciences Research Council UK (EPSRC/RCUK) (Grant Number: EP/P018165/1- Newton Fund) is gratefully acknowledged. The comments from the anonymous reviewers in improving this article are highly appreciated.

References

Afonja T. 2017. Kernel functions. [accessed 2019 March 12]. https://towardsdatascience.com/kernel-function-6f1d2be6091.

Belgiu M, Drăguț L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens. 114:24-31. doi:10.1016/j.isprsjprs.2016.01.011.

Bittencourt H, Clarke RT. 2003. Use of classification and regression trees (CART) to classify remotely-sensed digital images. In: IGARSS 2003. Proceedings of the IEEE Trans Geosci Remote Sens Symposium 2003. Toulouse, France, July 21–25: IEEE IGARSS. p. 3751–3753.

Brid RS. 2018. Decision trees. Medium; [accessed 2019 January 20]. https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb.

Bro-Jørgensen J, Brown ME, Pettorelli N. 2008. Using the satellite-derived normalized difference vegetation index (NDVI) to explain ranging patterns in a lek-breeding antelope: the importance of scale. Oecolgia. 158(1):177-182. doi:10.1007/s00442-008-1121-z

Calbury. 2016. Object-based Classification: Classification and Regression Tree (CART). [Accessed 2019 March 3]. wiki.landscapetoolbox.org/doku.php/remote_sensing_methods:classification_and_regression_tree_cart.

Camacho A, Correa CV, Arguello H. 2019. An analysis of spectral variability in hyperspectral imagery: a case study of stressed oil palm detection in Colombia. Int J Remote Sens. 40(19):7603-7623. doi:10.1080/01431161.2019.1595210.

Chander G, Markham BL, Helder DL. 2009. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sens Environ. 113(5):893-903. doi:10.1016/j.rse.2009.01.007.

Charters LJ, Aplin P, Marston CG, Padfield R, Rengasamy N, Bin Dahalan MP, Evers S. 2019. Peat swamp forest conservation withstands pervasive land conversion to oil palm plantation in North Selangor, Malaysia. Int J Remote Sens.40(19):1-30. doi:10.1080/01431161.2019.1574996.

Cheng Y, Yu L, Cracknell AP, Gong P. 2016. Oil palm mapping using Landsat and PALSAR: A case study in Malaysia. Int J Remote Sens. 37(22):5431-5442. doi:10.1080/01431161.2016.1241448.

Cheng Y, Yu L, Xu Y, Lu H, Cracknell AP, Kanniah K, Gong P. 2018. Mapping oil palm extent in Malaysia using ALOS-2 PALSAR-2 data. Int J Remote Sens. 39(2):432-452. doi:10.1080/01431161.2017.1387309.

Chong KL, Kanniah KD, Pohl C, Tan KP. 2017. A review of remote sensing applications for oil palm studies. Geo-spatial Information Science. 20(2):184-200. doi:10.1080/10095020.2017.1337317.

De Alban, J, Connette G, Oswald P, Webb E. (2018). Combined Landsat and L-band SAR data improves land cover classification and change detection in dynamic tropical landscapes. Remote Sensing. 10(2):306. doi:10.3390/rs10020306.

Dong J, Xiao X, Menarguez MA, Zhang G, Qin Y, Thau D, Biradar C, Moore B. 2016. Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine. Remote Sens Environ. 185:142-154. doi:10.1016/j.rse.2016.02.016.

Duro DC, Franklin SE, Dubé, MG. 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sens Environ. 118:259-272. doi:10.1016/j.rse.2011.11.020.

Fahmi Z, Samah BA, Abdullah H. 2013. Paddy industry and paddy farmers well-being: a success recipe for agriculture industry in Malaysia. Asian Soc Sci. 9(3):177-181. doi:10.5539/ass.v9n3p177.

Fawcett D, Azlan B, Hill TC, Kho LK, Bennie J, Anderson K. 2019. Unmanned aerial vehicle (UAV) derived structure-from-motion photogrammetry point clouds for oil palm (Elaeis guineensis) canopy segmentation and height estimation. Int J Remote Sens. 40(19):1-23. doi:10.1080/01431161.2019.1591651.

Fitzherbert EB, Struebig MJ, Morel A, Danielsen F, Brühl CA, Donald PF, Phalan B. 2008. How will oil palm expansion affect biodiversity?. Trends Ecol Evol. 23(10):538-545. doi:10.1016/j.tree.2008.06.012.

Gambo J, Shafri HZM, Shaharum NSN, Abidin FAZ, Rahman MTA. 2018. Monitoring and Predicting Land Use-Land Cover (LULC) Changes Within and Around Krau Wildlife Reserve (KWR) Protected Area in Malaysia Using Multi-Temporal Landsat Data. Geoplanning: journal of geomatics and planning. 5(1):23-44. doi:10.14710/geoplanning.5.1.17-34.

Gislason PO, Benediktsson JA, Sveinsson JR. 2006. Random forests for land cover classification. Pattern Recognit Lett. 27(4):294-300. doi:10.1016/j.patrec.2005.08.011.

Glinskis EA, Gutiérrez-Vélez VH. 2019. Quantifying and understanding land cover changes by large and small oil palm expansion regimes in the Peruvian Amazon. Land Use Policy. 80:95-106. doi:10.1016/j.landusepol.2018.09.032.

Goldblatt R, You W, Hanson G, Khandelwal A. 2016. Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine. Remote sensing. 8(8):634. doi:10.3390/rs8080634.

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens Environ. 202:18-27. doi:/10.1016/j.rse.2017.06.031.

Gutiérrez-Vélez VH, DeFries R. 2013. Annual multi-resolution detection of land cover conversion to oil palm in the Peruvian Amazon. Remote Sens Environ. 129:154-167. doi:10.1016/j.rse.2012.10.033.

Gupta O, Das AJ, Hellerstein J, Raskar R. 2018. Machine learning approaches for large scale classification of produce. Sci Rep. 8(1):5226-5233.

Ivanovic B. 2016. Cross-validation and decision trees. [accessed 2019 January 15]. https://www.cs.utoronto.ca/~fidler/teaching/2015/slides/CSC411/tutorial3_CrossVal-DTs.pdf.

Jiang L, Wang W, Yang X, Xie N, Cheng Y. 2010. Classification methods of remote sensing image based on decision tree technologies. Paper presented at CCTA 2010. International Conference on Computer and Computing Technologies in Agriculture; Oct 20–25; Nanchang, China.

Joshi N, Baumann M, Ehammer A, Fensholt R, Grogan K, Hostert P, Jepsen MR, Kuemmerie T, Meyfroidt P, Mitchard ETA, Reiche J, Ryan CM, Waske B. (2016). A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. Remote Sensing. 8(1):70. doi:10.3390/rs8010070.

Jusoff K, Pathan M. 2009. Mapping of individual oil palm trees using airborne hyperspectral sensing: an overview. Applied physics research. 1(1):15-30.

Lee JSH, Wich S, Widayati A, Koh LP. 2016. Detecting industrial oil palm plantations on Landsat images with Google Earth Engine. Remote sensing applications: society and environment. 4:219-224. doi:10.1016/j.rsase.2016.11.003.

Li L, Dong J, Tengku SN, Xiao X. 2015. Mapping oil palm plantations in Cameroon using PALSAR 50-m orthorectified mosaic images. Remote sensing. 7(2):1206-1224. doi:10.3390/rs70201206.

Li W, Fu H, Yu L, Cracknell A. (2016). Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. Remote Sensing. 9(1):22. doi:https://doi.org/10.3390/rs9010022.

Mahidin MU. 2018. Selected Agricultural Indicators Malaysia. Malaysia: Department of Statistics Malaysia; [accessed 2019 March 15]. https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=72&bul_id=UjYxeDNkZ0xOUjhFeHpna20wUUJOUT09&menu_id=Z0VTZGU1UHBUT1VJMFlpaXRRR0xpdz09.

Maxwell AE, Warner TA, Fang F. 2018. Implementation of machine-learning classification in remote sensing: An applied review. Int J Remote Sens. 39(9):2784-2817. doi:10.1080/01431161.2018.1433343.

Miettinen J, Shi C, Tan WJ, Liew SC. 2012. 2010 land cover map of insular Southeast Asia in 250-m spatial resolution. Pattern Recognit Lett. 3(1):11-20. doi:10.1080/01431161.2010.526971.

Molnar C. 2016. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 1st ed. Christoph Molnar. [accessed 2019 January 20]. https://christophm.github.io/interpretable-ml-book/tree.html.

Morel AC, Fisher JB, 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. Int J Remote Sens. 33(11):3614-3639. doi:10.1080/01431161.2011.631949.

Nambiappan B, Ismail A, Hashim N, Ismail N, Shahari DN, Idris NAN, Omar N, Salleh KM, Hassan NAM, Din AK. 2018. Malaysia: 100 years of resilient palm oil economic performance. J. Oil Palm Res. 30(1):13-25. doi:10.21894/jopr.2018.0014.

Ng WPQ, Lam HL, Ng FY, Kamal M, Lim JHE. 2012. Waste-to-wealth: green potential from palm biomass in Malaysia. J Clean Prod. 34:57-65. doi:10.1016/j.jclepro.2012.04.004.

Noi PT, Kappas M. 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors. 18(1):1-20. doi:10.3390/s18010018.

Nooni IK, Duker AA, Van Duren I, Addae-Wireko L, Osei Jnr EM. 2014. Support vector machine to map oil palm in a heterogeneous environment. Int J Remote Sens. 35(13):4778-4794. doi:10.1080/01431161.2014.930201.

Oliphant AJ, Thenkabail PS, Teluguntla P, Xiong J, Gumma MK, Congalton RG, Yadav K. 2019. Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud. Int J Appl Earth Obs Geoinf. 81:110-124. doi:10.1016/j.jag.2018.11.014.

Pal M. 2005. Random forest classifier for remote sensing classification. Int J Remote Sens. 26(1):217-222. doi:10.1080/01431160412331269698.

Pal M, Maxwell AE, Warner TA. 2013. Kernel-based extreme learning machine for remote-sensing image classification. Pattern Recognit Lett. 4(9):853-862. doi:10.1080/2150704X.2013.805279.

Patel NN, Angiuli E, Gamba P, Gaughan A, Lisini G, Stevens FR, Tatem AJ, Trianni G. 2015. Multitemporal settlement and population mapping from Landsat using Google Earth Engine. Int J Appl Earth Obs Geoinf. 35:199-208. doi:10.1016/j.jag.2014.09.005.

Patel S. 2017. Chapter 2: SVM (Support Vector Machine) — Theory. [accessed 2019 March 10]. https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

Roy DP, Wulder MA, Loveland TR, Woodcock CE, Allen RG, Anderson MC, Helder D, Irons JR, Johnson DM, Kennedy R, et al. 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sens Environ. 145:154-172. doi:10.1016/j.rse.2014.02.001.

Shafri HZM, Hamdan N. 2009. Hyperspectral imagery for mapping disease infection in oil palm plantationusing vegetation indices and red edge techniques. Am J Appl Sci. 6(6):1031-1035.

Shafri HZM, Anuar MI, Seman IA, Noor NM. 2011. Spectral discrimination of healthy and Ganoderma-infected oil palms from hyperspectral data. Int J Remote Sens. 32(22):7111-7129. doi:10.1080/01431161.2010.519003.

Shahar FM. 2016 March 18. Heat wave emergency if temperature exceeds 40 degrees Celsius for more than 7 days. New Straits Times. [accessed 2019 January 20]. https://www.nst.com.my/news/2016/03/133534/heat-wave-emergency-if-temperature-exceeds-40-degrees-celcius-more-7-days.

Shaharum NSN, Shafri HZM, Gambo J, Abidin FAZ. 2018. Mapping of Krau Wildlife Reserve (KWR) protected area using Landsat 8 and supervised classification algorithms. Remote Sensing Applications: Society and Environment. 10:24-35. doi:10.1016/j.rsase.2018.01.002.

Shaharum NSN, Shafri HZM, Ghani WAWAK, Samsatli S, Prince HM, Yusuf B, Hamud AM. 2019. Mapping the spatial distribution and changes of oil palm land cover using an open source cloud-based mapping platform. Int J Remote Sens. 40(19):1-18. doi:10.1080/01431161.2019.1597311.

Shelestov A, Lavreniuk M, Kussul N, Novikov A, Skakun S. 2017. Exploring Google earth engine platform for Big Data Processing: Classification of multi-temporal satellite imagery for crop mapping. Front Earth Sci. 5(17):1-10. doi:10.3389/feart.2017.00017.

Shuit SH, Tan KT, Lee KT, Kamaruddin AH. 2009. Oil palm biomass as a sustainable energy source: A Malaysian case study. Energy. 34(9):1225-1235. doi:10.1016/j.energy.2009.05.008.

Sidhu N, Pebesma E, Câmara G. 2018. Using Google Earth Engine to detect land cover change: Singapore as a use case. Eur J Remote Sens. 51(1):486-500. doi:10.1080/22797254.2018.1451782.

Thenkabail PS, Stucky N, Griscom BW, Ashton MS, 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. Int J Remote Sens. 25(23):5447-5472. doi:10.1080/01431160412331291279.

Yu X, Hyyppä J, Litkey P, Kaartinen H, Vastaranta M, Holopainen M. 2017. Single-sensor solution to tree species classification using multispectral airborne laser scanning. Remote Sensing. 9(2):108-123. doi:10.3390/rs9020108.

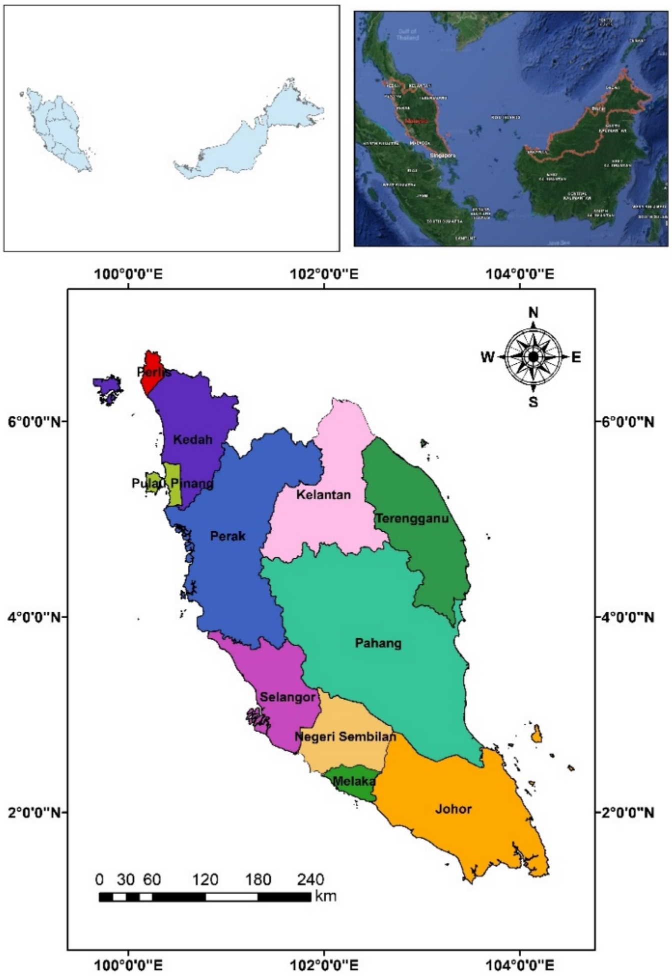

Figure 1. Location of the study area: Peninsular Malaysia.

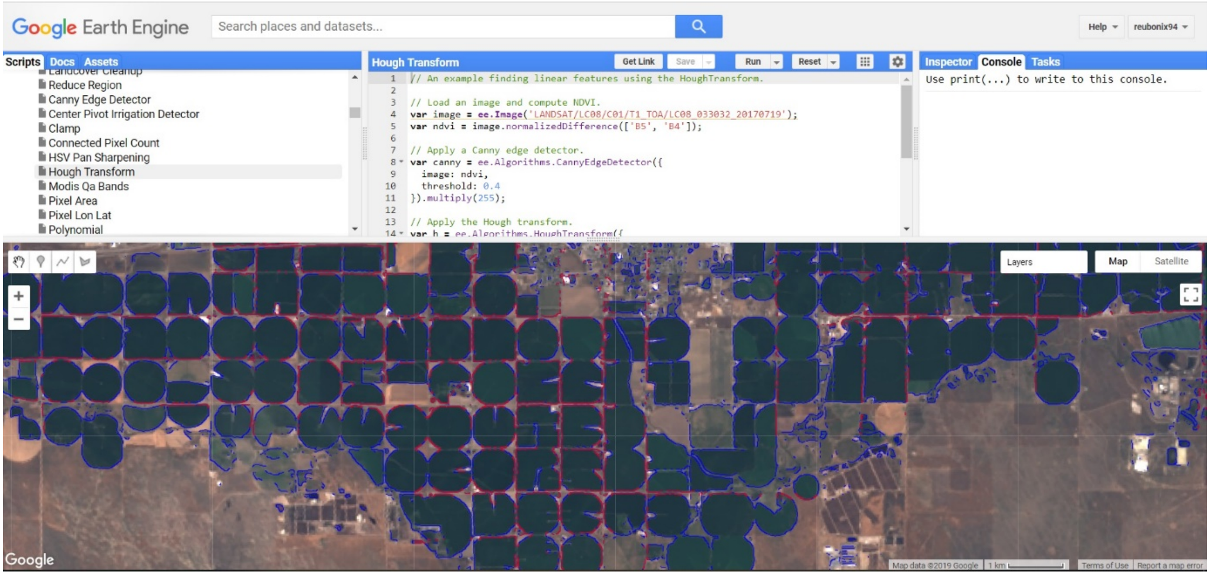

Figure 2. The Earth Engine Javascript API.

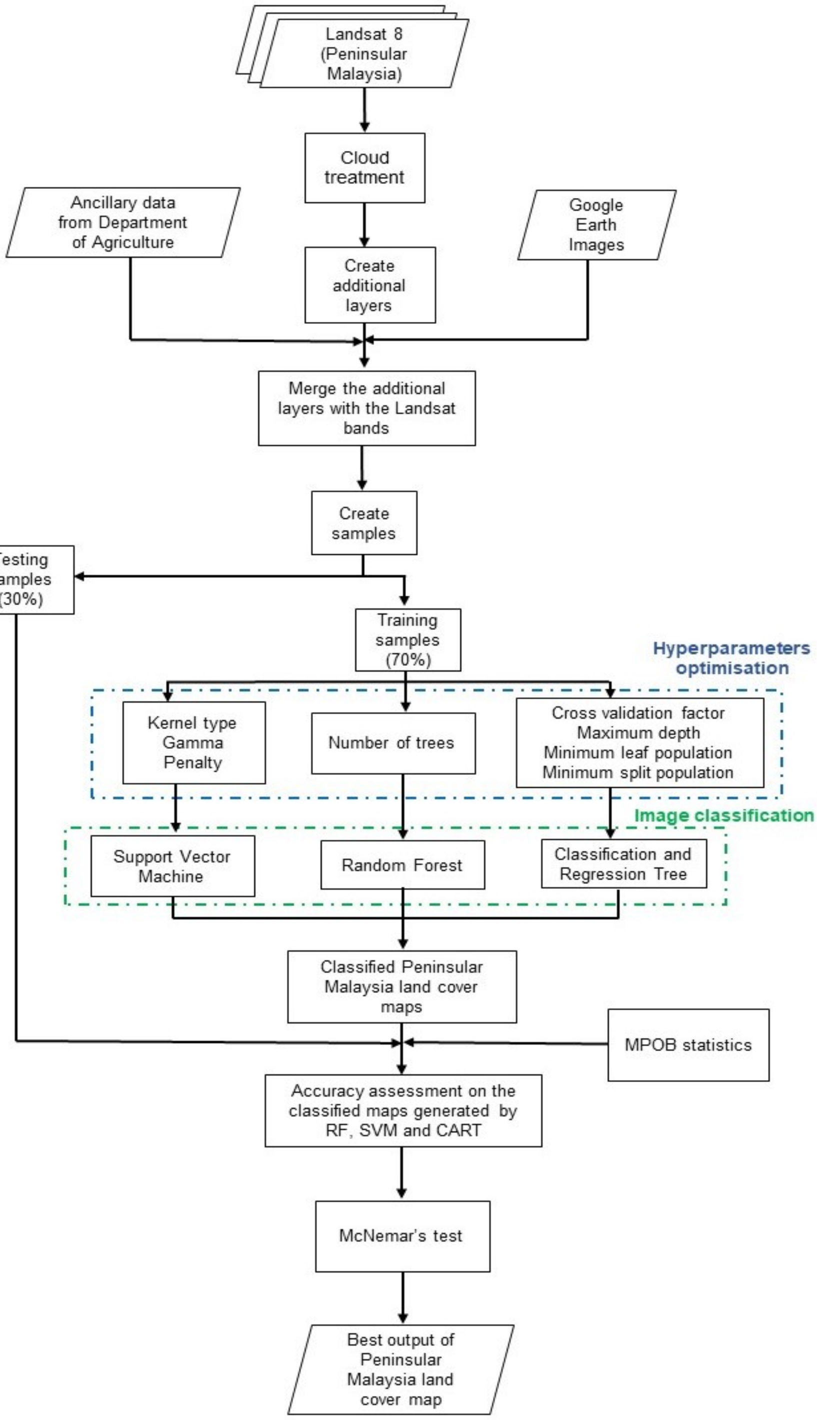

Figure 3. Methodological steps conducted for this study.

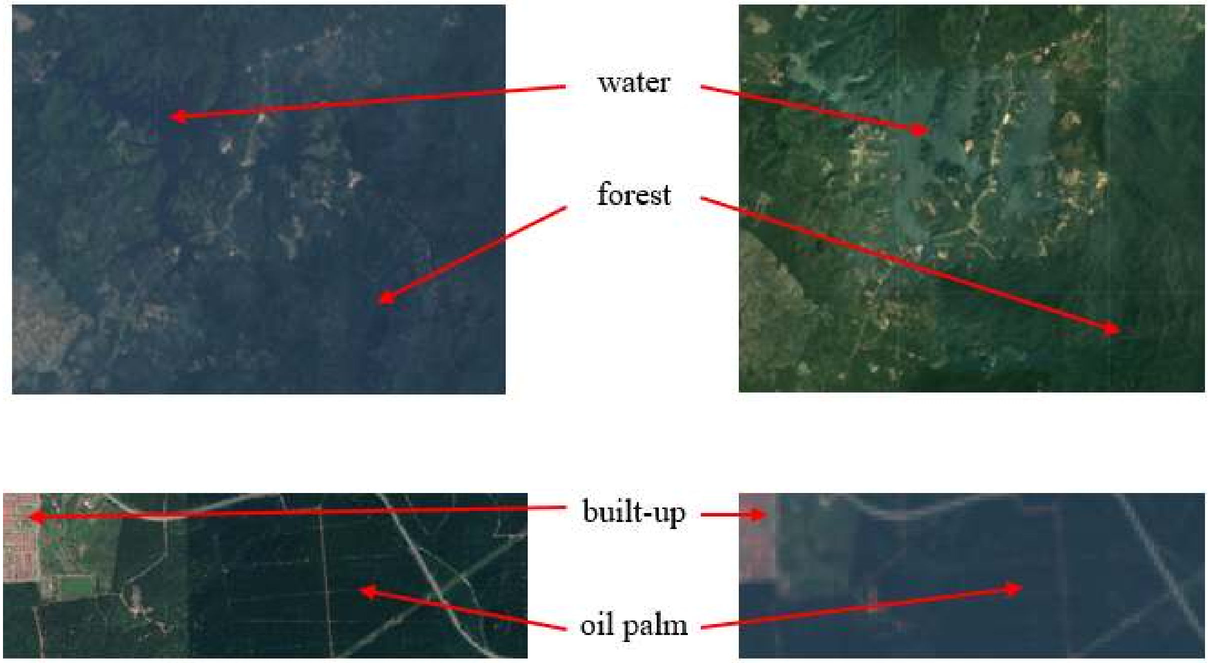

Figure 4. (a) High-resolution Google Earth image, (b) Landsat 8 image.

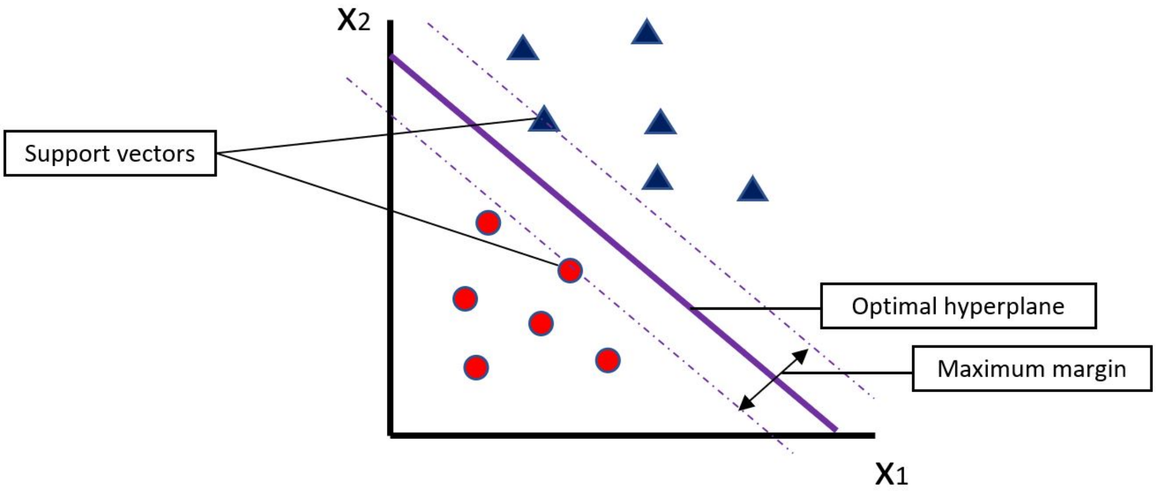

Figure 5. Optimal hyperplane identification in SVM.

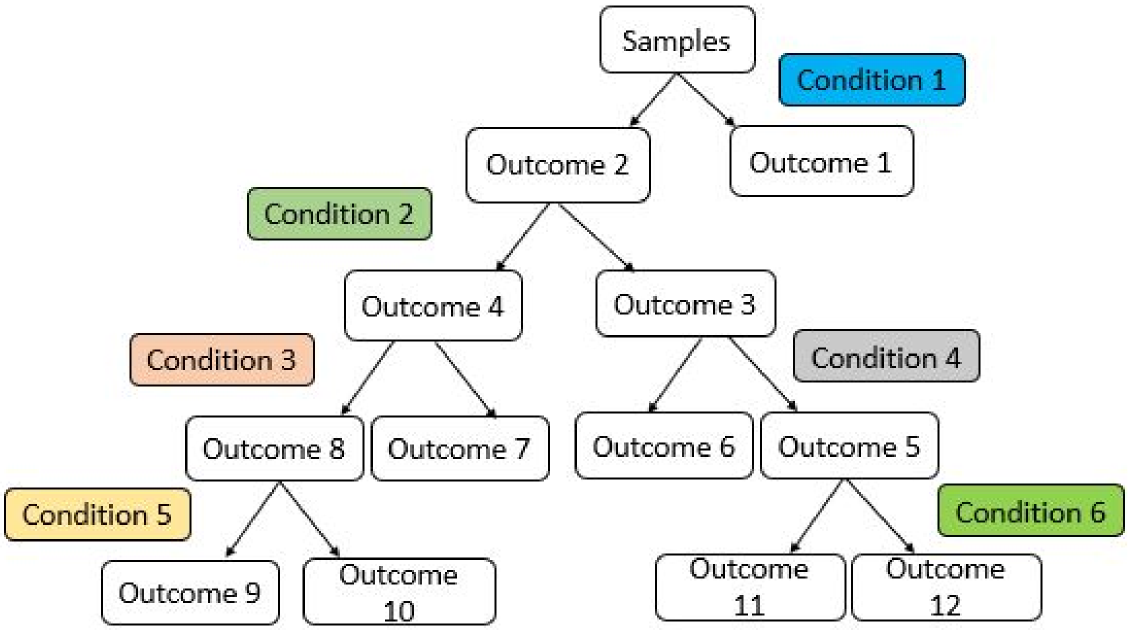

Figure 6. The division of the tree in CART.

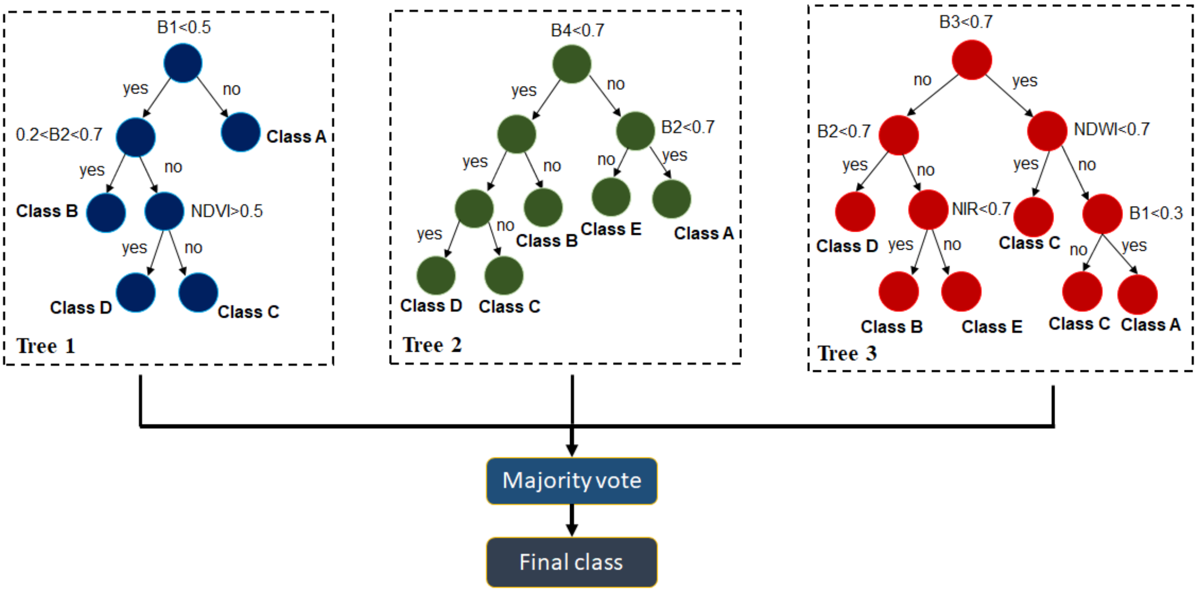

Figure 7. Example of trees ensemble in the RF structure.

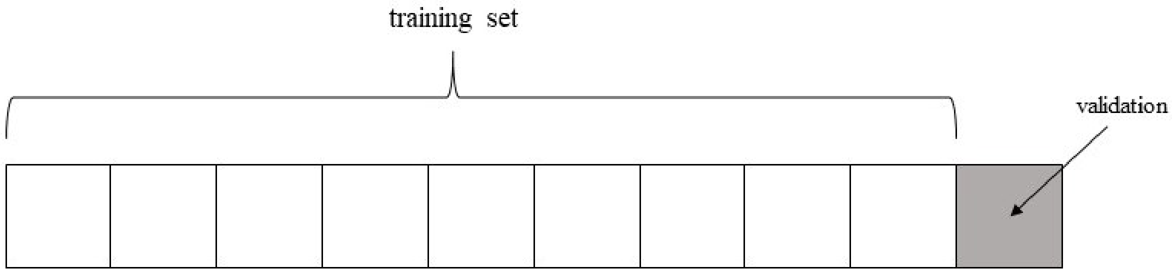

Figure 8. Subsamples in cross validation.

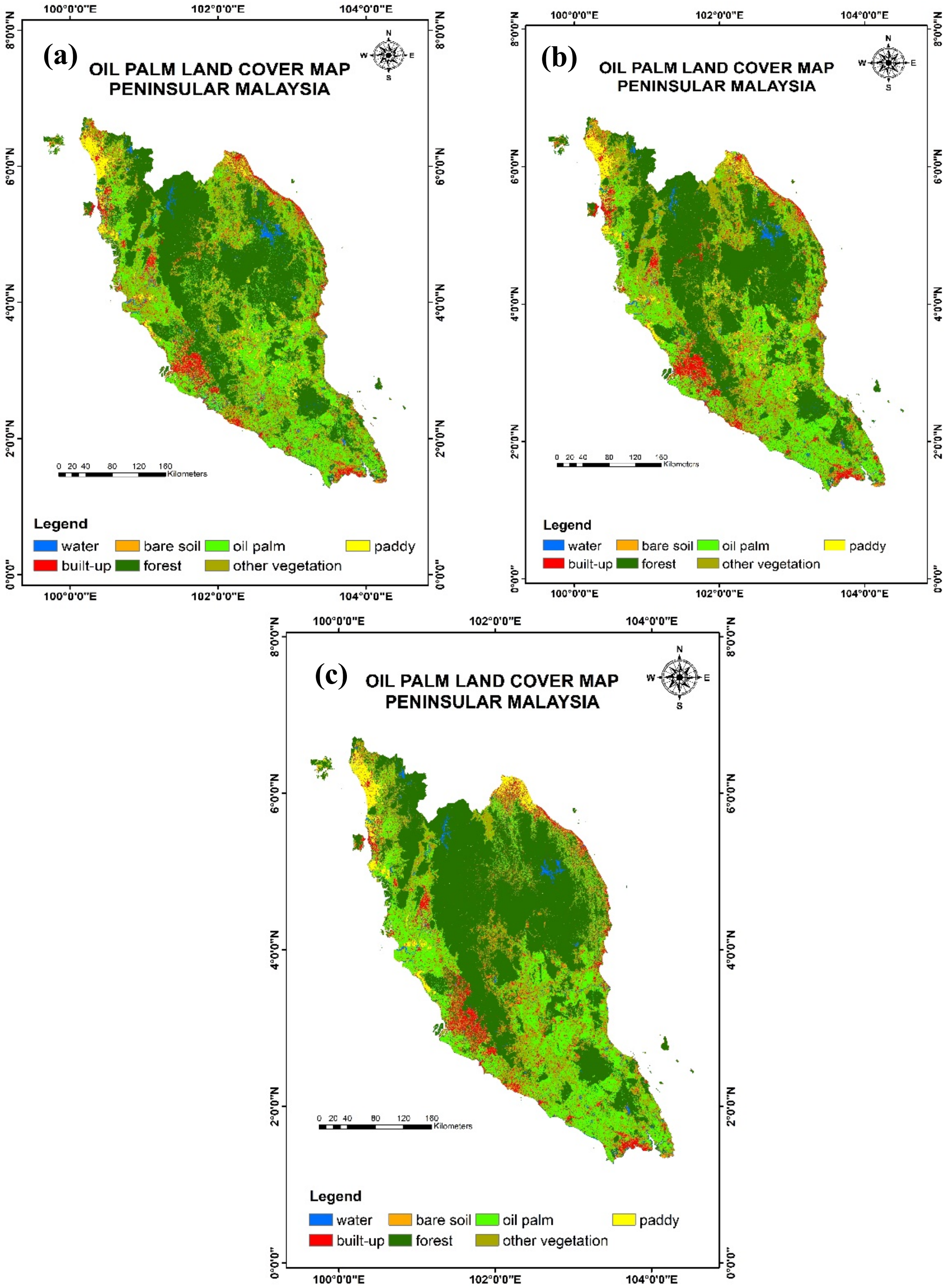

Figure 9. Classified oil palm land cover maps of Peninsular Malaysia, (a) CART, (b) RF and (c) SVM.

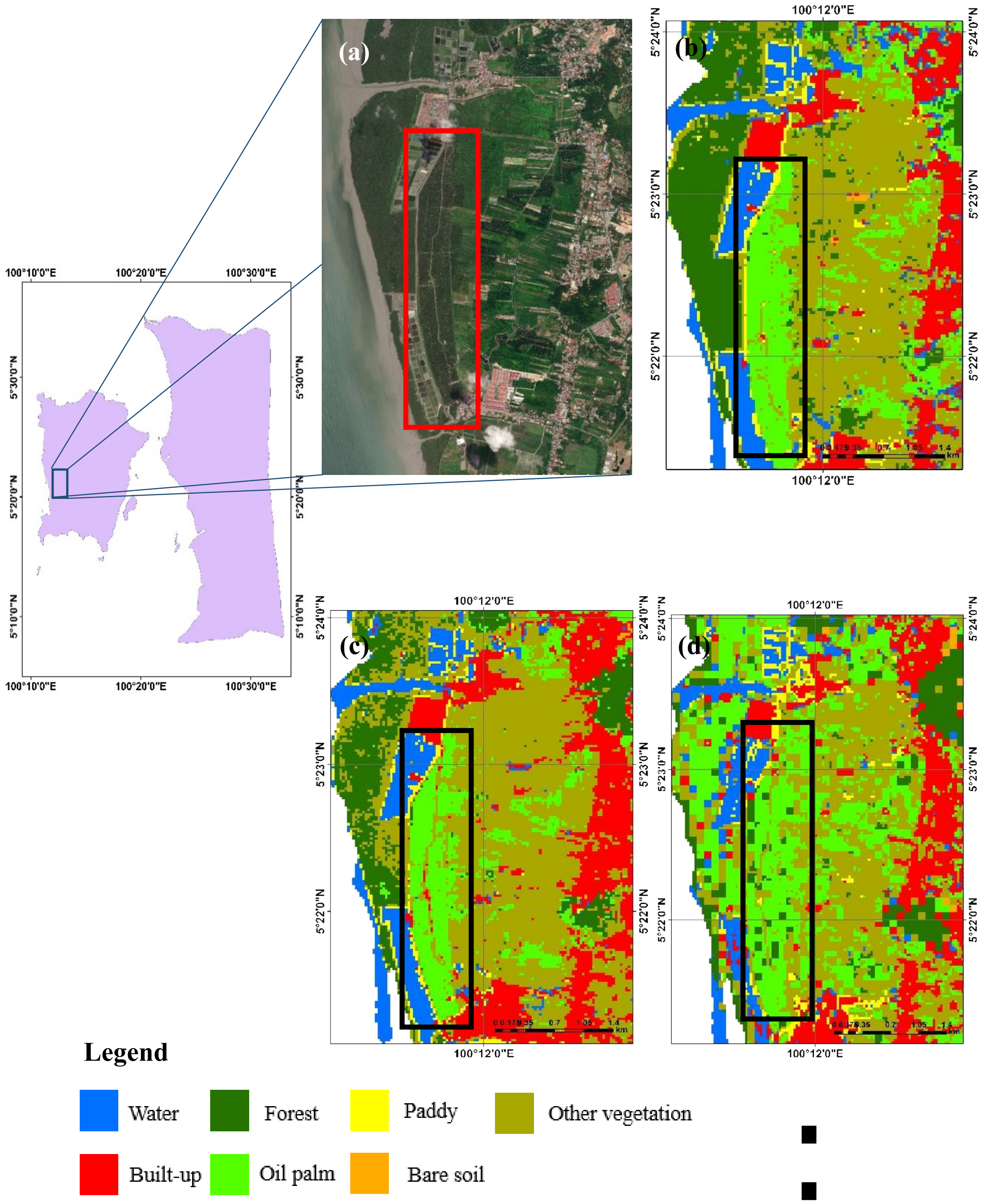

Figure 10. (a) High-resolution Google Earth image, (b) CART, (c) RF and (d) SVM.

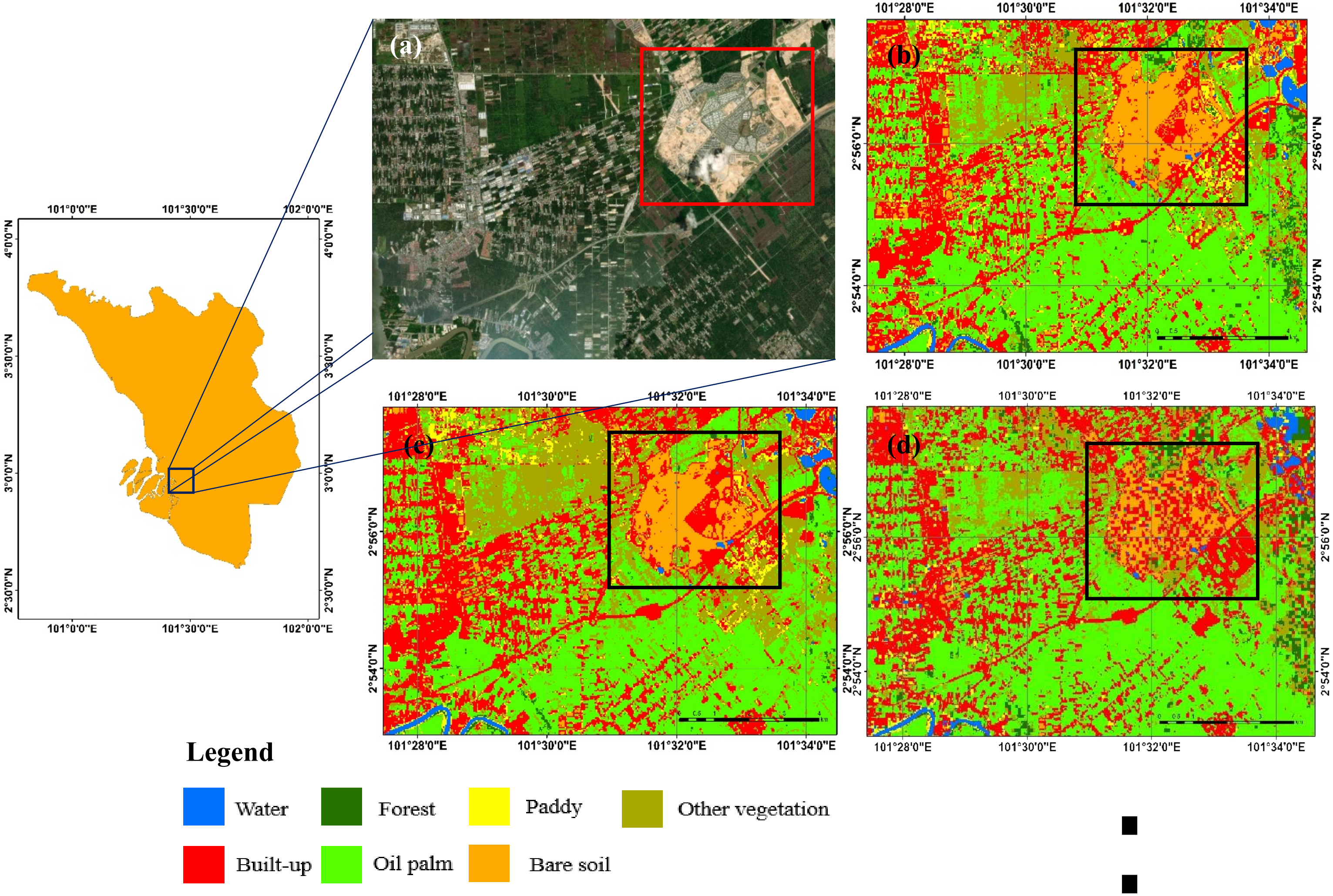

Figure 11. (a) High-resolution Google Earth image, (b) CART, (c) RF and (d) SVM.

Table 1. Information of the Landsat 8 bands.

Name	Description	Pixel size (m)	Wavelength (µm)
Band 1	Coastal aerosol	30	0.435 - 0.451
Band 2	Blue	30	0.452 - 0.512
Band 3	Green	30	0.533 - 0.590
Band 4	Red	30	0.636 - 0.673
Band 5	Near Infrared	30	0.851 - 0.879
Band 6	Short-wave Infrared 1	30	1.566 - 1.651
Band 7	Short-wave Infrared 2	30	2.107 - 2.294

Table 2. Additional layer to be included for classification.

Name	**Formula**	**Reference/Source**
NDVI	$\frac{NIR - Red}{NIR + Red}$	(Bannari et al., 1995; Maselli, 2004)
NDWI	$\frac{Green - NIR}{Green + NIR}$	(Xu et al., 2010)
Blue Red	$Blue - Red$	(Murray et al., 2018)
Blue Green	$Blue - Green$	

Table 3. Hyperparameters involved.

Algorithm	**Hyperparameter**
SVM	Kernel type = Radial Basis Function Gamma = 0.7 Penalty value = 10
CART	Cross validation factor = 5 Max depth = 10 Minimum leaf population = 5 Minimum split population = 10
RF	Number of trees = 30

Table 4. Overall, producer's and user's accuracies for oil palm class of each state and Peninsular Malaysia.

State		Johor	Kedah	Kelantan	Melaka	Negeri Sembilan	Pahang	Pulau Pinang	Perak	Perlis	Selangor	Terengganu	Peninsular Malaysia
RF	OA (%)	89.23	87.85	86.06	85.16	77.57	80.84	89.74	91.30	86.75	89.88	87.10	86.50
	PA (%)	84.62	100.00	89.66	92.00	68.75	80.49	93.10	81.82	85.71	92.45	87.50	86.92
	UA (%)	89.19	84.44	86.67	74.19	75.86	70.21	96.43	90.00	75.00	87.50	80.77	82.75
CART	OA (%)	82.74	86.74	73.94	87.50	78.04	76.64	80.13	82.61	69.88	78.75	83.87	80.08
	PA (%)	84.62	92.11	75.86	96.00	71.88	85.37	65.52	81.82	85.71	73.58	91.67	82.19
	UA (%)	76.74	85.37	73.33	80.00	58.97	70.00	65.52	81.82	50.00	88.64	59.46	71.82
SVM	OA (%)	89.38	97.35	98.18	88.28	88.32	81.28	96.15	97.10	97.59	97.62	93.55	93.16
	PA (%)	89.74	97.22	100.00	92.00	87.50	89.19	96.55	90.91	100.00	98.11	87.50	93.52
	UA (%)	85.37	97.22	93.55	82.14	80.00	76.74	96.55	90.91	100.00	96.30	91.30	90.01

Note: OA: Overall accuracy (7 classes); PA: Producer's accuracy (oil palm); UA: User's accuracy (oil palm)

Table 5. Oil palm area produced by RF, CART and SVM in comparison with MPOB.

State	Oil palm area (ha)						
	MPOB	RF		CART		SVM	
		Classified	Difference	Classified	Difference	Classified	Difference
Johor	748860	799142	50282	752133	3273	782282	33422
Kedah	87538	147744	60206	157801	70263	170060	82522
Kelantan	158310	126177	-32133	183388	25078	106969	-51341
Melaka	57372	45768	-11604	42186	-15186	46021	-11351
Negeri Sembilan	184815	184325	-490	195733	10918	194311	9496
Pahang	741495	720745	-20750	802325	60830	717739	-23756
Pulau Pinang	13563	13146	-417	16039	2476	16572	3009
Perak	406469	392518	-13951	445041	38572	528448	121979
Perlis	660	1779	1119	3760	3100	4789	4129
Selangor	137783	196807	59024	195375	57592	194506	56723
Terengganu	171548	167136	-4412	211977	40429	162737	-8811

Table 6. Contingency table.

	Test 2 (positive)	**Test 2 (negative)**	**Row total**
Test 1 (positive)	a	b	a + b
Test 1 (negative)	c	d	c + d
Column total	a + c	b + d	n

Table 7. McNemar’s test result.

Algorithm 1	**Algorithm 2**	**p-value**
SVM	RF	0.28
SVM	CART	0.00
RF	CART	0.00

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

- The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name	**Affiliation**
Nur Shafira Nisa Shaharum	Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia.
Helmi Zulhaidi Mohd Shafri	Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia. Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia.
Wan Azlina Wan Ab Karim Ghani	Department of Chemical and Environmental Engineering/Sustainable Process Engineering Research Centre (SPERC), Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia.
Sheila Samsatli	Department of Chemical Engineering, University of Bath, Claverton Down, BA2 7AY, United Kingdom.
Mohammed Mustafa Abdulrahman Al-Habshi	Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia.
Badronnisa Yusuf	Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia.

Ethical Statement for Remote Sensing Applications: Society and Environment

I testify on behalf of all co-authors that our article submitted to Solid State Ionics – Diffusion and Reactions:

Title: Oil Palm Mapping Over Peninsular Malaysia Using Google Earth Engine and Machine Learning Algorithms

All authors:

Nur Shafira Nisa Shaharum, Helmi Zulhaidi Mohd Shafri, Wan Azlina Wan Ab Karim Ghani, Sheila Samsatli, Mohammed Mustafa Abdulrahman Al-Habshi and Badronnisa Yusuf.

1) this material has not been published in whole or in part elsewhere;
2) the manuscript is not currently being considered for publication in another journal;
3) all authors have been personally and actively involved in substantive work leading to the manuscript, and will hold themselves jointly and individually responsible for its content.

Date: 3rd Oct 2019

Corresponding author's signature:

6 January 2020

Editor
Remote Sensing Applications: Society and Environment

Dear Sir,

<u>SUBMISSION OF AN UPDATED MANUSCRIPT FOR CONSIDERATION OF PUBLICATION IN THE REMOTE SENSING APPLICATIONS: SOCIETY AND ENVIRONMENT JOURNAL</u>

With reference to the matter as stated above, I would like to submit a manuscript entitled **"Oil Palm Mapping Over Peninsular Malaysia Using Google Earth Engine and Machine Learning Algorithms"** for consideration of publication in the Remote Sensing Applications: Society and Environment Journal. This paper requires improvements after being revised with the decision of minor correction.

We thank the editors and reviewers for the comments to help improve the quality of the manuscript. We have highlighted (using yellow highlighting with red font color) the changes made in our manuscript based on the comments.

The details on the response to the reviewers are given in the correction table for your reference and I hope it will receive a proper evaluation from the Journal reviewers and editors.

Best regards,

Helmi Zulhaidi Mohd Shafri
Dept. of Civil Engineering
UPM