IMAGE CLASSIFICATION FOR MAPPING OIL PALM DISTRIBUTION VIA SUPPORT VECTOR MACHINE USING SCIKIT-LEARN MODULE

N. S. N. Shaharum[1], H. Z. M. Shafri[1, 2, *], W. A. W. A. K. Ghani[3], S. Samsatli[4], B. Yusuf [1], M. M. A. Al-Habshi[1], H. M. Prince[1]

[1]Department of Civil Engineering, Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia
[2]Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia – helmi@upm.edu.my
[3]Department of Chemical and Environmental Engineering/Sustainable Process Engineering Research Centre (SPERC), Faculty of Engineering, Universiti Putra Malaysia, 43400, UPM Serdang, Selangor, Malaysia
[4] Department of Chemical Engineering, University of Bath, Claverton Down, BA2 7AY, United Kingdom.

KEY WORDS: Landsat, oil palm, Python, remote sensing, Scikit-learn, support vector machine.

ABSTRACT:

The world has been alarmed with the global warming effects. Global warming has been a distress towards the environment, thus shorten the Earth's lifespan. It is a challenging task to reduce the global warming effects in a short period, knowing that the human population is increasing along with the electricity and energy demand. In order to reduce the effects, renewable energy is presented as an alternative method to produce energy in a way that will not harm the environment. Oil palm is one of the agricultural crops that produces huge amount of biomass which can be processed and used as a renewable energy source. In 2016, Malaysia has reported over 5 million hectares of land were covered by oil palm plantations. Placing Malaysia as the second largest country of oil palm producer in the world has given it an advantage to produce renewable energy source. However, there is a need to monitor the sustainability of oil palm plantations in Malaysia via effective mapping approaches. This study utilised two different platforms (open source and commercial) using a machine learning algorithm namely Support Vector Machine (SVM) to perform oil palm mapping. An open source Python programming-based technique utilising Scikit-learn module was performed to map the oil palm distribution and the result produced had an overall accuracy of 91.39%. To support and validate the efficiency of the Python programming-based image classification, a commercial remote sensing software (ENVI) was used and compared by implementing the same SVM algorithm and the result showed an overall accuracy of 98.21%.

1. INTRODUCTION

Today, energy crisis has become a serious issue especially for developing countries. Subsequently, the energy demand is increasing as their population is growing (Mekhilef et al., 2011; Ong et al., 2011). Mekhilef et al. (2011) stated that an alternative way needs to be carried out in order to replace the uses of fossil fuels to generate energy. This is because fossil fuels can no longer withstand in the near future due to the impacts towards the environment. Malaysia is blessed with humid and tropical climate which directly puts Malaysia as the second largest oil palm producer in the world.

Oil palm is one of the major vegetable oils and it has been widely used worldwide. Furthermore, oil palm is one of the biomass resources that can be used as a source of energy (Loh, 2017). Bio-diesel extracted from palm oil is biodegradable, safe, and non-toxic, thus makes it suitable to be used as a renewable energy source. In Malaysia, the oil palm plantations had an increment over the years and over 5 million hectares of oil palm area was reported in 2016 (Table 1). Therefore, Malaysia has the potential not only to produce renewable energy source, but also to be used as cooking oil and other food products (Aziz et al., 2011; Umar et al., 2014; Mba et al., 2015). However, it is a big challenge to manage a huge area of oil palm plantations especially when there are many things need to be done and properly planned. Therefore, a proper strategy with suitable and adequate information are essential in order to have an effective plan management. Due to the huge amount of data required, remote sensing offers an effective method to help in a way such data can be obtained.

Year	Oil palm area for Malaysia (ha)
2013	5,229,739
2014	5,392,235
2015	5,642,943
2016	5,737,985

Table 1. Oil palm plantation for Malaysia (MPOB, 2017)

Remote sensing is the science of acquiring information without making a direct contact with the object. It has been used in numerous number of fields and disciplines such as agriculture, urban areas, geography, and land surveying (Joshi et al., 2016; Razali et al., 2016; Norman et al., 2017). Furthermore, remote sensing is not only capable of acquiring data in inaccessible area, but also can obtain huge amount of data in a very short time. Furthermore, remote sensing is possible to collect data from various sensors (active and passive) and platforms including ground-based, aerial-based, and satellite-based. Then, the collected data are normally being processed and classified using suitable remote sensing or Geographic Information System (GIS) software such as ERDAS (ERDAS, Inc), ENVI (ITT Visual Information Solutions, Boulder, CO, USA), ArcMap GIS software, and SNAP (Sentinel Application Platform). Basically, software provides tools for image processing which includes image calibration, classification, and accuracy assessment. In other word, software serves as a platform to perform image analysis and map making using various approach and algorithms. Several algorithms available for image classifications are supervised and unsupervised

algorithms including Random Forest (RF), Support Vector Machine (SVM), and Maximum Likelihood Classifier (MLC). Li et al. (2015) conducted a study on mapping oil palm in Cameroon using Palsar 50m orthorectified mosaicked images. The study had utilised SVM, Decision Tree (DT), and K-Mean algorithms for image classifications. Above all the algorithms mentioned, SVM was found as the ideal algorithm for oil palm mapping. Another study on oil palm mapping conducted by Lee et al. (2016) had utilised Landsat data obtained from Google Earth Engine (GEE). GEE is a cloud-based platform that allows the user to perform image analysis including data acquisition and image analysis. The platform uses Javascript that requires the user to code in order to obtain the data from the cloud server. Other than cloud-based platform, coding via programming languages can also be used for remote sensing analysis. A popular language known as Python has been widely used for image classifications, machine learning analysis, and deep learning approaches.

Python is one of the well-known programming languages that is widely used in various fields including data analysis and predictions (Predegosa et al., 2011; Li et al., 2017). There are many libraries available in Python that can be used to perform image analysis including image pre-processing, image classification, and also to produce land use land cover (LULC) map. In addition to that, Python-based image classification allows the user to tune the hyperparameters within the algorithm. The flexibility of Python programming allows the user to choose and design the procedure based on the user's needs. Due to the effectiveness of machine learning on multispectral data as mentioned by Shafri (2017), this study has used a supervised machine learning algorithm that was imported from Scikit-learn module. Due to its great performance in previous studies conducted by Peña et al. (2014), Nooni et al. (2014), and Gilbertson et al. (2017), SVM was chosen to be used to map the oil palm distribution. Owing to the versatility of Python programming language in providing number of libraries, this study was conducted to assess the capability of the programming-based using Python version 3.5 to map the oil palm distribution via SVM algorithm. Then, the result obtained will be compared with a well-known commercial remote sensing software, ENVI (Goetz, 2009).

2. STUDY AREA AND SATELLITE DATA

This study was conducted within Selangor area. Selangor is one of the states where its land is covered with oil palm plantations (MPOB, 2017). To test the Python programming-based approach for image classification, a pilot study was conducted in Sepang, which is located at the southern part of Selangor. The area was chosen due to its coverage that consists of different features and furthermore, the area has the least amount of clouds. An open source data obtained from Landsat 8 satellite was used in this study. The data with the least cloud cover acquired on 29th March 2016 was used and the data comes with 11 bands including Multispectral, Panchromatic, and Thermal bands. Figure 1 showed Landsat 8 image of the study area with the combination of band 4, 3, and 2 (true colour).

In order to increase the quality of the image, a pan-sharpening technique was applied using the panchromatic band (Gilbertson et al., 2017; Shaharum et al., 2018). This technique was conducted to increase the spatial resolution from 30m to 15m. The capability of near-infrared band has proved to be a success in differentiating green vegetations from other features which would be helpful for oil palm detection (Candiago et al., 2015; Roy et al., 2016). Besides utilising panchromatic band for image enhancement, only multispectral bands were utilised in this study for image classification.

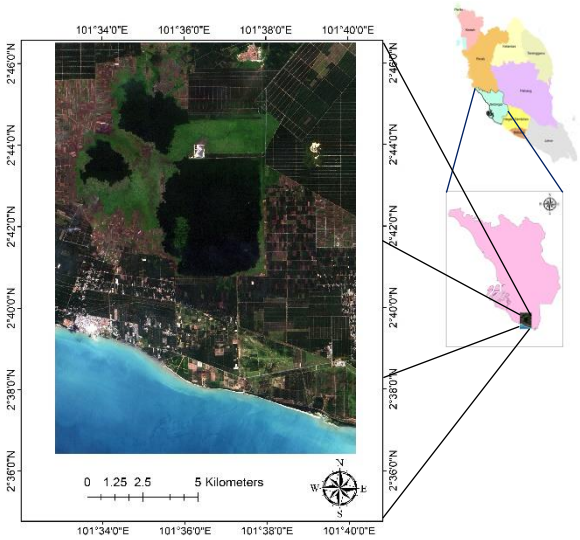

Figure 1. Study area

3. METHODOLOGY

The satellite image taken from Landsat 8 was downloaded from https://earthexplorer.usgs.gov/ and the image was chosen based on the minimum cloud cover. The obtained image was pre-processed in ENVI version 5.3 (ITT Visual Information Solutions, Boulder, CO, USA) and which later being exported as a tiff file format.

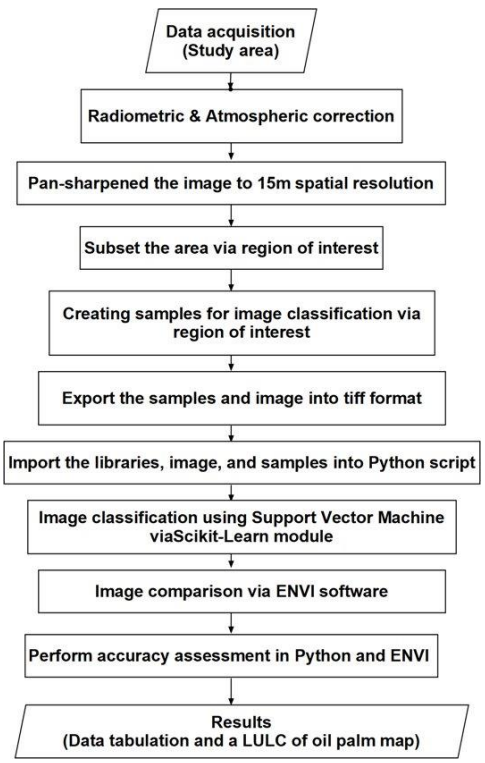

Figure 2. Flow chart for the work flow

3.1 Image Pre-Processing

The downloaded image is a raw image and it needs to be corrected. Atmospheric and radiometric corrections were applied on the image in ENVI software by converting the Digital Number (DN) to reflectance value. Each pixel consists of different reflectance value depending on its feature. Later, these values were then being assessed by the algorithm to classify the features based on the assigned training and testing samples.

3.2 Development of Training Samples

The samples were created via Region of Interest (ROI) in ENVI based on the selected features using square polygons. Four classes were created namely oil palm, built-up/road, non-oil palm, and water. Each class was assigned with a certain number of ROIs and colour. The selection of the ROIs was done based on the high-resolution image from Google Earth. Then, the ROI samples created were exported to tiff file format where it can be used to classify the image via Scikit-learn SVM module in Python.

3.3 Libraries in Python for Image Classification

Several libraries such as GDAL, Numpy, Scikit-Learn, and Matplotlib were imported into the Python script. Each library has its own functions and capabilities which made them possible to be used for image classification (Predegosa et al., 2011). The pre-processed image and the samples were imported using GDAL. To perform the image classification in Python, the samples should be assigned accordingly to the georeferenced image. Therefore, to ensure that the samples are placed correctly according to the assigned feature, the samples were geo-coordinated by using the satellite image as the reference.

3.3.1 Support Vector Machine

SVM is an advanced machine learning algorithm that works by separating the support vectors at maximum distance by using a hyperplane (Müller et al., 1997; Mountrakis et al., 2011; Tehrany et al., 2015). It can work well even with the limited number of samples. A number of kernels are available in SVM and Radial Basis Function (RBF) was chosen to classify the image as the results from previous studies showed that RBF is the most superior kernel (Foody and Mathur, 2004; Bekios-Calfa et al., 2011). The common parameters presented in RBF were gamma and penalty and these parameters were tuned in order to produce the best result.

3.3.2 Accuracy Assessment

The samples were divided into 70/30 ratio whereby 70% taken from the whole samples was used to classify the image. Then, the other 30% was used to validate the output produced in a form of a classified image. The assessment was done using a train-test-split module in Python that was imported from the Scikit-learn module.

3.4 Ground Truthing

The ground truthing was conducted based on the available high-resolution image from Google Earth image and a reference from the LULC map provided by the Department of Agriculture (DOA). These available references were not only being used as an aid in producing the samples, but also to validate the outputs produced.

4. RESULTS AND DISCUSSION

The parameters of SVM were adjusted and the best output produced was used to represent the oil palm distributions for the area. To measure the capabilities of utilising Python programming-based, the result produced was compared by classifying the image using the same algorithm and parameters in a commercial software, ENVI.

4.1 Classification of Oil Palm

Four classes (water, non-oil palm, built-up, and oil palm) were classified and the classified image produced in Python and ENVI were exported to a tiff file format as shown in Figure 3 and Figure 4 respectively. The area consists of numerous number of vegetations and other features including ponds, buildings, and oil palms. Other than oil palm, all vegetations and trees were classified as the non-oil palm feature.

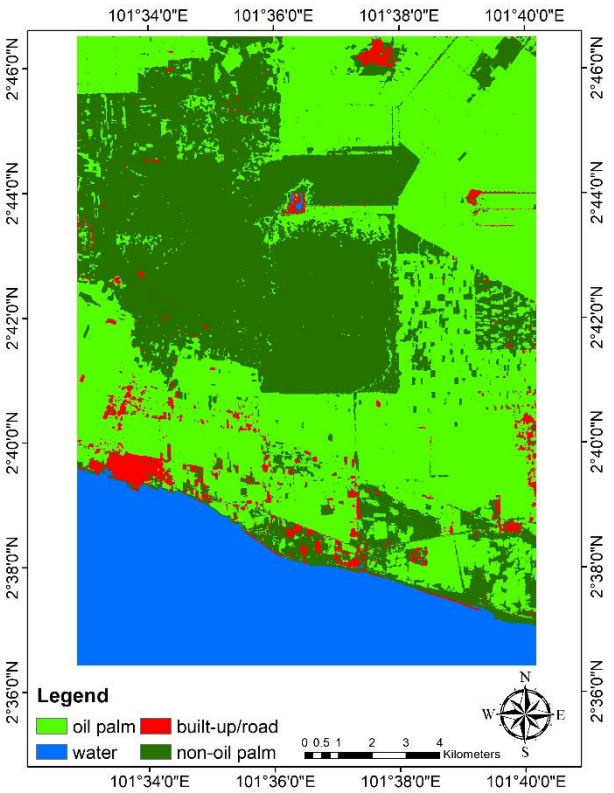

Figure 3. SVM classified image using Python

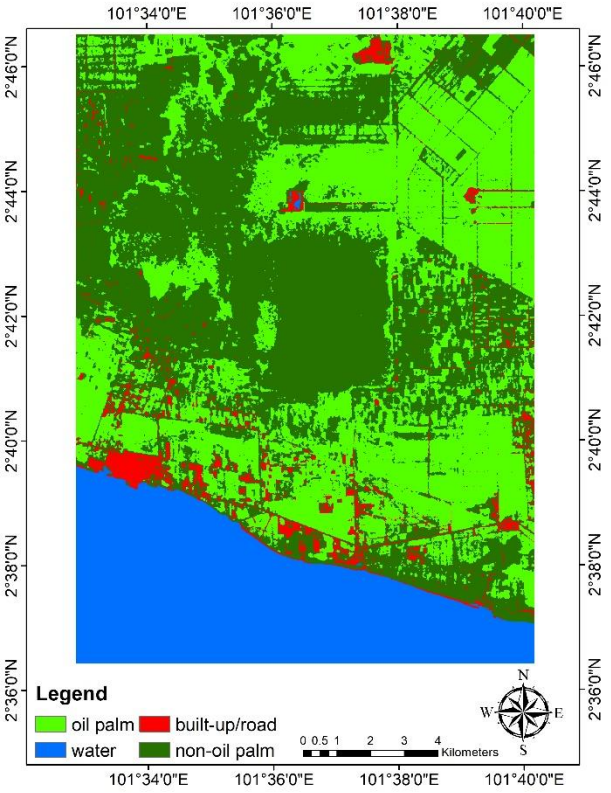

Figure 4. SVM classified image using ENVI

4.2 Discussion

The overall accuracy (OA) produced in Python and ENVI were 91.39% and 98.21% respectively. Though Figure 3 and 4 showed almost similar results, the OA produced in ENVI was higher than the OA produced in Python. The confusion matrix for both results produced by Python and ENVI were shown in Table 2 and 3 respectively.

Class	Op	W	Bu/R	N-op
Op	610	0	6	3
W	0	180	0	0
Bu/R	67	0	46	15
N-op	58	0	3	849

Table 2. Confusion matrix produced in Python

Class	W	Op	N-op	Bu/R
W	78	0	0	0
Op	0	58	3	0
N-op	0	1	67	0
Bu/R	0	0	0	17

Table 3. Confusion matrix produced in ENVI

Op = Oil palm, N-op = Non-oil palm, W = Water, Bu/R = Built-up/Road

Table 2 showed some misclassifications occurred between built-up/road and oil palm features. Then, a bit confusion was found between non-oil palm and oil palm. On the other hand, Table 3 showed less confusion between oil palm and non-oil palm features. However, the visualisation for the classified maps showed almost similar results for the class of oil palm and non-oil palm. Even though OA produced in Python was lower than the OA produced in ENVI, the output produced in Python was said to comply better to the reality.

4.3 Conclusion

Python programming-based utilising Scikit-learn to perform SVM classification managed to produce a reasonable output. It can identify the oil palm distributions similar to the software-based technique though the OA produced in Python was lower than the OA produced in ENVI. On top of that, the time taken for the SVM classification applied in Python was shorter than the commercial software-based SVM classification. This method can later be tested on a larger area for further assessment. In a nutshell, the performance of Python is convincing (based on the benchmarking with the industry-standard software e.g. ENVI) and provides a cost-effective and innovative alternative as it is open source and free.

4.4 Future direction

There are few methods that can be done in order to assess and measure the accuracy of the outputs produced. Depending to only one source might not be sufficient to evaluate the accuracy of the algorithms performed as the OA produced alone does not define the precision of the output. Furthermore, besides SVM, Python programming provides other algorithms such as RF, Neural Network, and other machine learning algorithms which later can be tested on other satellite data with different sensor and spatial resolutions.

ACKNOWLEDGEMENTS

The author would like to thank UPM for their facilities and funding of this research. Apart from that, our heartfelt gratitude also goes to Engineering and Physical Sciences Research Council for their financial support through the BEFEW (Newton Fund) project (Grant No. EP/P018165/1).

REFERENCES

Aziz, M. A. A., Ali. S., Ghani, W. A. W. A. K., Saleh, M. A. M., Ahmadun, F., & Taufiq-Yap, Y. H. (2011). Hydrogen rich gas from oil palm biomass as a potential source of renewable energy in Malaysia. *Renewable and Sustainable Energy Reviews*, *15*(2), 1258-1270.

Bekios-Calfa, J., Buenaposada, J. M., & Baumela, L. (2011). Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(4), 858-864.

Candiago, S., Remondino, F., De Giglio, M., Dubbini, M., & Gattelli, M. (2015). Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. *Remote Sensing*, *7*(4), 4026-4047.

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(6), 1335-1343.

Gilbertson, J. K., Kemp, J., & Van Niekerk, A. (2017). Effect of pan-sharpening multi-temporal Landsat 8 imagery for crop type differentiation using different classification

techniques. *Computers and electronics in agriculture*, *134*, 151-159.

Goetz, A. F. (2009). Three decades of hyperspectral remote sensing of the Earth: A personal view. *Remote Sensing of Environment*, *113*(Sep), S5-S16.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jephsen, M. R., Kuemmerle, T., Meyfroidt. P., Mitchard, E. T. A., Reiche, J., Ryan, C. M., & 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.

Lee, J. S. H., Wich, S., Widayati, A., & Koh, L. P. (2016). Detecting industrial oil palm plantations on Landsat images with Googe Earth Engine. *Remote Sensing Applications: Society and Environment*, *4*(October), 219-224.

Li, W., Wu, G., Zhang, F., & Du, Q. (2017). Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(2), 844-853.

Loh, S. K. (2017). The potential of the Malaysian oil palm biomass as a renewable energy source. *Energy Conversion and Management*, *141*, 285-298.

Mba, O. I., Dumont, M. J., & Ngadi, M. (2015). Palm oil: Processing, characterization and utilization in the food industry–A review. *Food bioscience*, *10*(June), 26-41.

Mekhilef, S., Saidur, R., Safari, A., & Mustaffa, W. E. S. B. (2011). Biomass energy in Malaysia: current state and prospects. *Renewable and Sustainable Energy Reviews*, *15*(7), 3360-3370.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247-259.

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997, October). Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks* (pp. 999-1004). Springer, Berlin, Heidelberg.

MPOB (Malaysian Palm Oil Board), 2017, Overview of the Malaysian Oil Palm Industry 2016, accessed 28.06.2018.

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.

Norman, M., Shafri, H. Z. M., Pradhan, B., & Yusuf, B. (2017, July). Improved Building Roof Type Classification Using Correlation-Based Feature Selection and Gain Ratio Algorithms. In *Global Civil Engineering Conference* (pp. 863-873). Springer, Singapore.

Okoro, S. U., Schickhoff, U., Böhner, J., & Schneider, U. A. (2016). A novel approach in monitoring land-cover change in the tropics: oil palm cultivation in the Niger Delta, Nigeria. *DIE ERDE–Journal of the Geographical Society of Berlin*, *147*(1), 40-52.

Ong, H. C., Mahlia, T. M. I., & Masjuki, H. H. (2011). A review on energy scenario and sustainable energy in Malaysia. *Renewable and Sustainable Energy Reviews*, *15*(1), 639-647.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubour, V., Vanderplas, J., Passos, A., Cournapeau, D., Brusher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Peña, J. M., Gutiérrez, P. A., Hervás-Martínez, C., Six, J., Plant, R. E., & López-Granados, F. (2014). Object-based image classification of summer crops with machine learning methods. *Remote Sensing*, *6*(6), 5019-5041.

Razali, S. M., Atucha, A. A. M., Nuruddin, A. A., Hamid, H. A., & Shafri, H. Z. M. (2016). Monitoring vegetation drought using MODIS remote sensing indices for natural forest and plantation areas. *Journal of Spatial Science*, *61*(1), 157-172.

Roy, D. P., Kovalskyy, V., Zhang, H. K., Vermote, E. F., Yan, L., Kumar, S. S., & Egorov, A. (2016). Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sensing of Environment*, *185*(November), 57-70.

Shafri, H. Z. M. (2017, June). Machine Learning in Hyperspectral and Multispectral Remote Sensing Data Analysis. In *Artificial Intelligence Science And Technology-Proceedings Of The 2016 International Conference (Aist2016)* (p. 3).

Shaharum, N. S. N., Shafri, H. Z. M., Gambo, J., & Abidin, F. A. Z. (2018). Mapping of Krau Wildlife Reserve (KWR) protected area using Landsat 8 and supervised classification algorithms. *Remote Sensing Applications: Society and Environment*, *10*(April), 24-35.

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-basesd support vector machine model with different kernel types. *Catena*, *125*(February), 91-101.

Umar, M. S., Jennings, P., & Urmee, T. (2014). Sustainable electricity generation from oil palm biomass wastes in Malaysia: An industry survey. *Energy*, *67*(April), 496-505.

Revised August 2018