RANDOM FOREST CLASSIFICATION FOR PLANTATION OF RUBBER AREA USING DIGITAL NUMBER VERSUS REFLECTANCE VALUES

Tengku Noradilah Tengku Jalal[1], N .Rajkumar A/L V. Nagarethinam[1], Nur Hurriyatul Huda Abdullah Sani[1],Nurul Aishah Rahman[1]

ABSTRAK

Perangkaan getah merupakan petunjuk penting kepada kerajaan dalam merancang penggubalan dasar dan pembangunan industri getah di Malaysia. Walau bagaimanapun, Malaysia bergantung kepada pengumpulan data konvensional iaitu kerja lapangan yang sukar untuk memantau kawasan ladang getah. Selaras dengan kemajuan dan perkembangan teknologi, DOSM telah mula mengambil satu lagi lonjakan dengan meneroka Pemerhatian Bumi (EO) untuk berinovasi dan meningkatkan kualiti statistik rasmi. Oleh itu kajian ini bertujuan untuk menentukan kawasan tanaman getah menggunakan dua data input berbeza iaitu nombor digital (DN) dan nilai pantulan (RV). Dalam kajian ini, daerah Hulu Perak yang terletak di Perak telah dipilih untuk menilai prestasi input menggunakan model Pembelajaran Mesin (ML) Random Forest (RF) yang dicadangkan dalam mengenal pasti kawasan getah menggunakan imej satelit Sentinel-2A pada 15 Mac 2019. Keputusan menunjukkan bahawa tiada perbezaan dalam ketepatan model apabila input ditukar. Oleh itu, imej sentinel-2A dengan nilai nombor digital boleh digunakan terus untuk membina model klasifikasi.

Kata kunci: Ladang getah, Hutan Rawak (RF), Nombor Digital (DN), Nilai Pantulan (RV), Sentinel-2A

ABSTRACT

Rubber statistics is an important indicator to the government in planning its policy formulation and development of the rubber industry in Malaysia. However, Malaysia relies on conventional data collection i.e. field works which can be difficult to monitor the rubber plantation area. In line with the progress and development of technology, DOSM has begun to innovate and improve the quality of official statistics by exploring Earth Observation (EO). Therefore, this study aims to determine plantation of rubber area using two different inputs: digital number (DN) and reflectance values (RV). In this study, we assess the performance of the proposed Random Forest (RF) Machine Learning (ML) model in identifying rubber areas in Hulu Perak district, Perak using Sentinel-2A satellite images on March 15, 2019. The result demonstrates that the

---

[1]Tengku Noradilah Tengku Jalal is currently Assistant Director of of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia; [2]N.Rajkumar A/L Nagarethinam is currently Principle Assistant Director of Agricultural and Environmental Statistic Division, Department of Statistics Malaysia; [1]Nur Hurriyatul Huda Abdullah Sani is currently Senior Assistant Director of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia and [1]Nurul Aishah Rahman is currently Assistant Director of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia.

accuracy of the models does not change when the inputs are changed. This study found that Sentinel-2A images with digital number values can be used directly to build classification models for rubber plantation area with the same accuracy as models built using reflectance values. This finding has important implications for future rubber plantation mapping studies, as it reduces the need for time-consuming and expensive reflectance value pre-processing.

Keywords: Rubber plantation, Random Forest (RF), Digital Number (DN), Reflectance Values (RV), Sentinel-2A

1. INTRODUCTION

Rubber, along with oil palm, timber, cocoa, pepper, and kenaf, is an important commodity contributing significantly to national income. Natural rubber survey not only the glove, tyre, and tube industries, but it is also a substantial export for Malaysia, with exports totalling 48,589 metric tonnes in June 2021. Since its first statistics made available in 1965, rubber statistics have served as crucial indicators and inputs for research, government agencies, and policymakers in planning, monitoring, and evaluating the performance of Malaysia's rubber industry.

Statistics record that the total area planted with rubber in 1965 was 788.5 thousand hectares, with a tapped area of 542.3 thousand hectares producing 507.9 thousand tonnes of natural rubber. However, as the Malaysian economy shifted from primary industries towards manufacturing and focus on other commodities, rubber statistics continued to record a downward trend. In 2019, the planted area was only 95.4 thousand hectares with a tapped area of 43.6 thousand hectares, which produced 61.2 thousand tonnes of natural rubber production.

Monitoring plays an essential role in agricultural management and production (Fukatsu, T., & Nanseki, T., 2009). Through effective and intensive monitoring, producers are able to identify corrective and preventive steps to optimise input while maximising production. Traditionally, rubber tree monitoring is time-consuming and labour-intensive. The collection of ground data relies heavily on conventional monitoring methods. (Nguyen, T. T,et.al, 2020). Currently, the rubber statistics relies on conventional data collection i.e., field works which can be challenging to monitor the rubber plantation area. Rubber industry is currently facing various challenges such as unpredictable weather and price instability. Existing method of data collection and administrative records prone to be affected by error for various reasons. In addition, restricted control movement during COVID-19 pandemic has made it even more challenging for field data collection.

The demand for accurate, reliable and more frequently measured data is increasing. This led national statistical agencies to explore various technology for innovative and more efficient methods in collecting agricultural data. In recent years the application of remote sensing and satellite-based technologies has gained attention from various Statistical Agencies. Remote sensing is about acquiring images and information about the Earth's land and water surfaces from an overhead perspective by employing electromagnetic radiation where a distinction is made between reflected or emitted electromagnetic radiation recorded by the sensors (Campbell and Wynne, 2011).

Scientists use satellite remote sensors to measure and map the density of green vegetation over the earth. A general definition of Remote Sensing is “the science and technology by which the characteristics of objects of interest can be identified, measured or analysed without direct contact” (JARS, 1993).

In alignment with technological progress, Satellite Imagery (SI) and Machine Learning (ML) offers great opportunities and potential in advancing and modernising rubber data collection methods. This study aims to determine plantation of rubber area using two different inputs data namely digital number (DN) and reflectance values (RV). In this study, we assess the performance of the proposed Random Forest (RF) Machine Learning (ML) model in identifying rubber areas in Hulu Perak district, Perak using Sentinel-2A satellite images on March 15, 2019.

A remote sensing system utilises detectors to sense energy reflected or emitted from the earth's surface, which might be modified by the intervening atmosphere. The sensor can be on a satellite, aircraft, or drone (Schowengerdt, 2006). The sensor turns the energy into a voltage, which an analogue to digital converter turns into a single integer value called the DN. A digital detector can store the DN directly. The value can be displayed with an appropriate colour to build up an image of the region sensed by the system. The DN represents the energy sensed by the sensor in a particular part of the electromagnetic spectrum, emitted or reflected from a particular region.

Radiance is the amount of radiation coming from an area. Radiance includes radiation reflected from the surface, bounced in from neighbouring pixels, and reflected from clouds above the area of the pixel. Radiance is also affected by the source of the radiation, which for optical imagery is the sun. Looking at the spectrum of a radiance pixel, it will have the overall shape of the solar spectrum, which peaks at green wavelengths (about 500 nm). Meanwhile reflectance is the proportion of the radiation striking a surface to the radiation reflected off of it. Some materials can be identified by their reflectance spectra, so it is common to correct an image to RV as a first step toward locating or identifying features in an image (Shippert, 2017).

2. METHODOLOGY

This study focuses in Hulu Perak, the biggest district (6,613 km^2) in Perak, Peninsular Malaysia. It experiences a humid tropical climate over the year (Figure 1). This district is surrounded by Jeli and Gua Musang District to the east, Kuala Kangsar District to the south and Larut and Matang, Selama and Baling District to the west. Hulu Perak has been chosen as the study area because of its geographically diverse, with the economic activity based on agriculture such as the cultivation of rubber, palm oil and other agricultural activities, as well as eco-tourism.

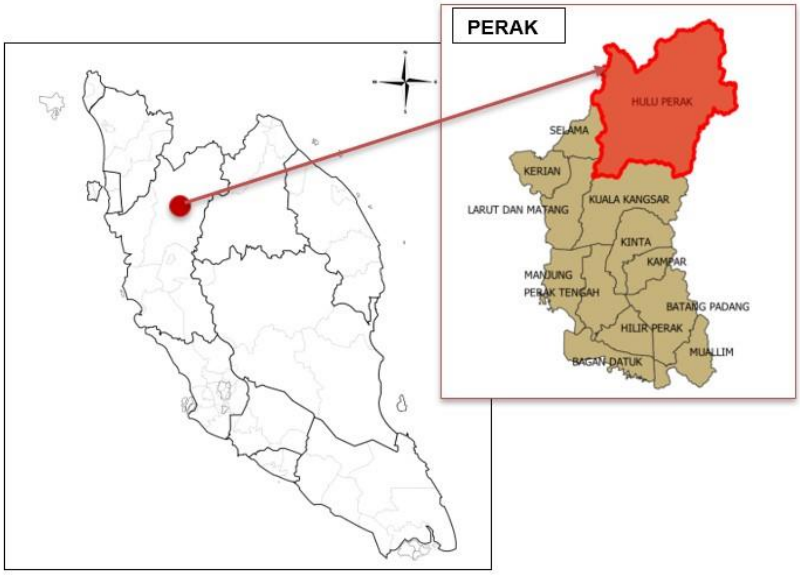

Figure 1: Study Area of Hulu Perak District, Perak

The four types of resolution that are important in remote sensing are spatial resolution, which refers to the fineness of details visible in an image, temporal resolution which denotes the time it takes for a sensor to return to a precise area or location, spectral resolution refers to the specific wavelength intervals whereas a sensor can record, and radiometric resolution determines accuracy of the sensor can distinguish between different levels of reflection (Rotairo et al., 2019). In this study, spectral resolutions from Sentinel-2A satellite image data captured on March 15, 2019, were acquired and utilized for rubber and other class image classification. These images were downloaded using Semi-Automatic Classification Plugin (SCP), a plugin available in Quantum Geography Information System (QGIS). The SCP provides tools to download, preprocessing and postprocessing images (Congedo, 2016). The Sentinel-2A images have different spatial resolutions from 10 to 60 metres with 13 spectral bands. Furthermore, the Department of Statistics Malaysia (DOSM) and the Town and Country Planning Malaysia Department (PLANMalaysia) provided geospatial data was utilized in this study for sample selection and validation.

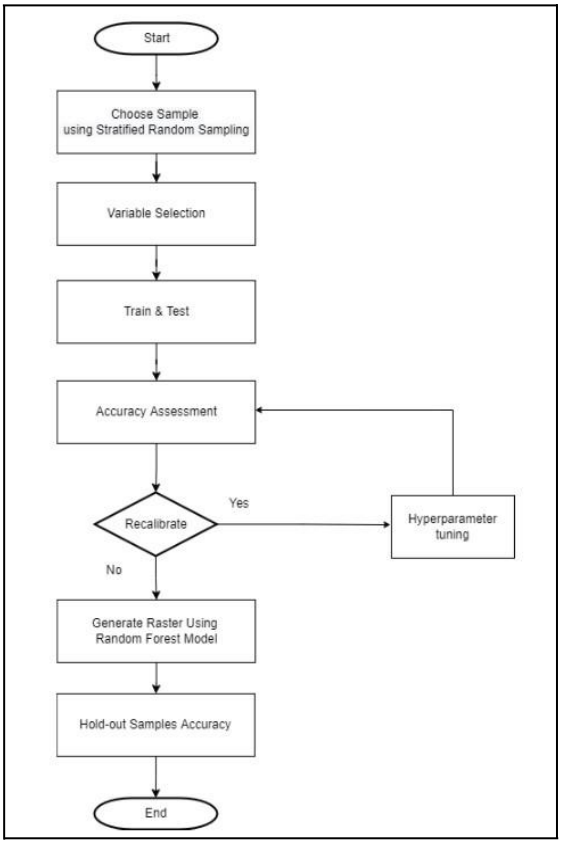

Figure 2: Flowchart of Random Forest Classification for Rubber and Other Class

Figure 2 shows the flow of the study. Sample for seven distinct classes, namely water, build-up area, rubber, palm oil, other vegetation, forest and bare lands were assigned. Homogeneous points for all 7 classes were selected using stratified random sampling and were extracted using Python Scripts in QGIS (Moyroud, Portet, 2018). The data were split into train (17,285), test (7,409) and validation (6,175) samples. Test data sets were used to determine unbiased evaluation of a model fit on the training dataset, tuning model hyperparameters and perform feature selection. Validation dataset are used to assess the performance of a fully-specified classifier model (Speiser et al., 2019).

Initially, there are 4 bands with 10- metre resolution, 6 bands with 20- metre resolution, along with 3 additional derived indices (Normalised Difference Vegetation Index (NDVI), Normalised Difference Building Index (NBDI) and Enhanced Vegetation Index (EVI)) used for this study (Table 1). Bands with 20- metre resolutions were resampled to 10- metre resolution to standardise the resolution. Pearson product-moment correlation coefficient conducted to check for association of the variables. Variables with weak correlation and equal mean difference will be removed from the study.

Table 1: List of bands used in the study

Sentinel-2 Bands	Resolution (m)
Band 2 - Blue	10
Band 3 - Green	10
Band 4 - Red	10
Band 5 - Vegetation Red Edge (VRE1)	20
Band 6 - Vegetation Red Edge (VRE2)	20
Band 7 - Vegetation Red Edge (VRE3)	20
Band 8 - Near Infrared (NIR)	10
Band 8A - Vegetation Red Edge (VRE4)	20
Band 11 - Short-wave infrared (SWIR1)	20
Band 11 - Short-wave infrared (SWIR2)	20
Normalised Difference Vegetation Index (NDVI)	-
Normalise Difference Building Index (NBDI)	-
Enhanced Vegetation Index (EVI)	-

Model was built using RF Classifier, an ensemble-based learning algorithm which consists of many decision trees developed by Leo Brieman to reach a single result (Breiman, 1996). RF are simple to implement, fast in operation and have proven to be extremely successful in various domains (Caie, Dimitriou & Arandjelović, 2021). This model is robust to outliers and indifferent to nonlinear features. Default parameters were used to build the initial model and later assessed using Confusion Matrix. Confusion Matrix tables predicted classification against the actual classification and derived indicators such as model accuracy, precision, recall and F1-score.

The model is later further recalibrated and tweaked using hyperparameter tuning to improve the model accuracy. Parameters which define the model architecture are referred to as hyperparameters and thus, this process of searching for the ideal model architecture is referred to as hyperparameter tuning. Hyperparameter tuning is an essential part of controlling the behaviour of a machine learning model. If model hyperparameters are not tuned correctly the estimated model parameters would produce suboptimal results, as they do not minimize the loss function and resulting model making more errors. Besides applying hyperparameter tuning, classical approaches are used in variable selection using backward stepwise selection. It is demonstrated that optimal feature selection has a positive effect on the accuracy and efficiency (Rogers & Gunn, 2005).

The goal is to reduce the number of variables and obtain a predictor that can reduce the burden of data gatherings and improve model efficiency. Raster image for the best RF Model is being generated from Open Source Geospatial (OSGeo) resources using Python Scripts. Hold-out or validation accuracy is performed to ensure that the best model performs optimally and unbiased in the wild.

3. RESULTS

Extracted pixels for each studied class shows that rubber and palm oil have the closest mean and the same spectral characteristics. Forest also shows the same spectral characteristics but with lower mean. Spectral characteristics for Other vegetation have

higher mean for band blue, green, red, VRE2, SWIR1 and SWIR2 while, lower mean was observed in band VRE2, VRE3 and VRE4 as compared to rubber, palm oil and forest (Figure 3).

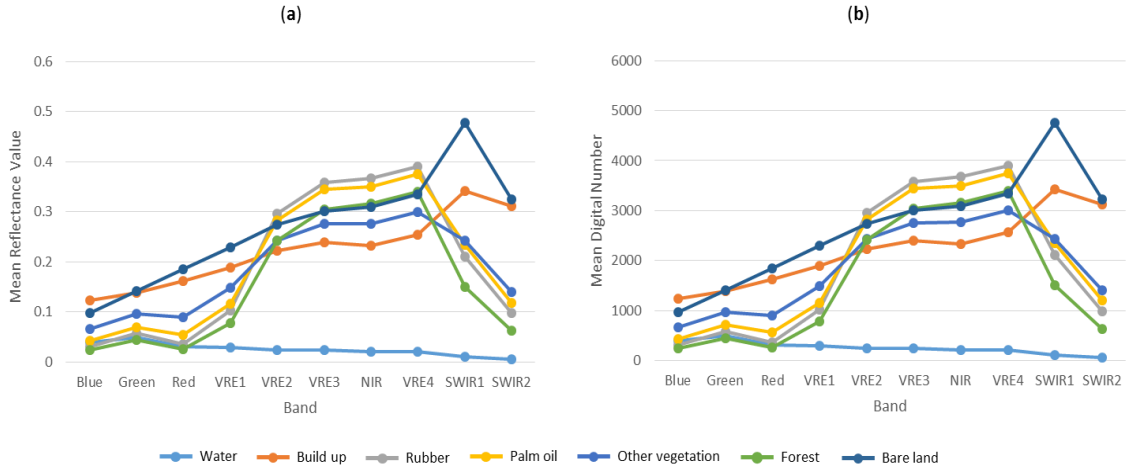

Figure 3: Comparison of the mean of different classes.
(a) Digital number values (b) Reflectance Values

However, after performing the Pearson product-moment correlation coefficient, EVI indicates a very weak association with the dependent variable (Table 2) and there was no statistically significant difference between the mean of EVI and other independent variables.

Table 2: Pearson product-moment correlation coefficient for EVI

		Variables											
		Blue	Green	Red	VRE1	VRE2	VRE3	NIR	VRE4	SWIR1	SWIR2	NDVI	NDBI
Input	DN	-0.015	-0.014	-0.013	-0.013	-0.010	-0.009	-0.008	-0.009	-0.011	-0.012	0.021	-0.009
	RV	-0.011	-0.011	-0.011	-0.011	-0.009	-0.008	-0.009	-0.007	-0.009	-0.010	0.004	-0.003

The backward stepwise selection output in Table 3 suggested reducing the number of variables to 7. Final models were run with only band blue, VRE1, VRE2, VRE3, NIR, VRE4 and SWIR2 as the independent variables.

Table 3: Backward stepwise selection output

Variables	p-value
Blue	< 0.0001
Green	0.025
Red	0.067
VRE1	< 0.0001
VRE2	< 0.0001
VRE3	< 0.0001
NIR	< 0.0001
VRE4	< 0.0001
SWIR1	0.092
SWIR2	< 0.0001
NDVI	0.025
NDBI	0.084

Table 4 shows the accuracy comparison of the model before hyperparameter tuning and variable selection for digital number and reflectance values. Overall accuracy of model with digital number input is 0.001% more accurate compared to reflectance values. F1-score for rubber classification also shows that the digital number score was 0.01 higher than the reflectance values.

Table 4: Comparison of rubber area accuracy assessment for digital number and reflectance values before hyperparameter tuning and variable selection

		Digital Number Values	Reflectance Values
Model	Overall Accuracy	0.962	0.961
Rubber	F1-Score	0.93	0.92

Table 5 shows the accuracy comparison of RF classification models between digital number and reflectance values for the final models. The confusion matrix indicates that there is a slight difference in overall accuracy between digital number and reflectance values with the result of 0.967% accuracy for digital number values and 0.968% accuracy for reflectance values. The result of Cohen's Kappa value which is 0.956 for digital number values and 0.954 for reflectance values shows that both inputs strongly help the model to successfully map all the classes. F1-score of rubber classification from the model stated that both input data gives the same accuracy score.

Table 5: Comparison of rubber area accuracy assessment for digital number and reflectance values for the final model

		Digital Number Values	Reflectance Values
Model	Overall Accuracy	0.967	0.968
	Cohen Kappa	0.956	0.954
Rubber	F1-Score	0.94	0.94

Figure 4 shows the spectral characteristics for the selected bands of the different classes. Rubber has similar spectral characteristics with palm oil and forest. Mean of digital number and reflectance value for rubber are slightly different to palm oil while forest has a lower mean.

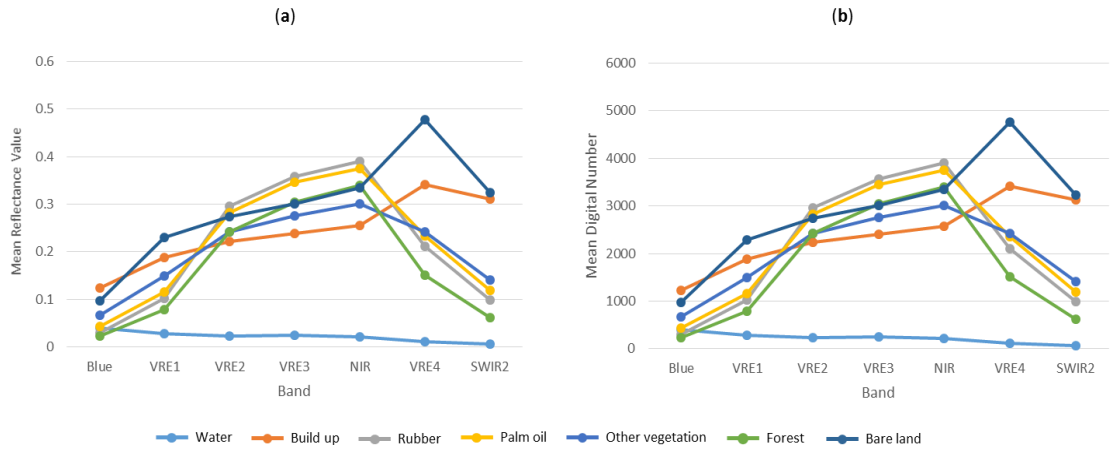

Figure 4: Comparison of the mean digital number and reflectance values of different classes.
(a) Digital number values (b) Reflectance Values

Thus, the raster of rubber classification result from RF are compared with rubber plantation land use provided by PLANMalaysia. The result of digital number and reflectance value shows a similar pattern with the data from PLANMalaysia (Figure 5). Mixed pixels effect may give some effect to the classification result since the spectral characteristics of rubber were similar to forest and palm oil.

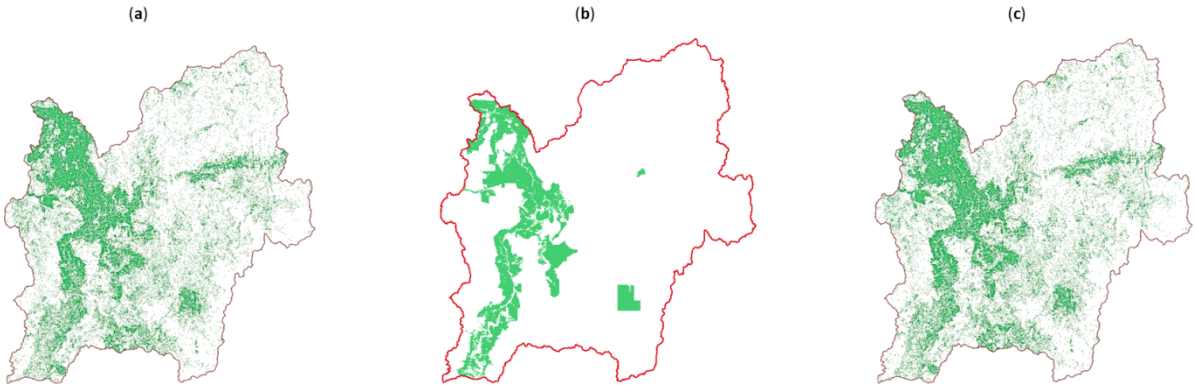

Figure 5: Comparison of the rubber distribution in Hulu Perak.
(a) RF classification based on digital number values (b) Rubber distribution stated by PLANMalaysia (c) RF classification based on reflectance values

4. CONCLUTION

The main aim of this study was to apply an RF based classification method with the difference input of sentinel-2 data (digital number and reflectance values) in determining the plantation of rubber area. This study demonstrates that there is no difference in accuracy of the models when the inputs were changed. The spectral characteristics demonstrate that rubber shares similarities with both to forest and palm oil. Thus, the mixed pixels effect may affect the result.

REFERENCES

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Caie, P. D., Dimitriou, N., & Arandjelović, O. (2021). Precision Medicine in digital pathology via Image Analysis and machine learning. *Artificial Intelligence and Deep Learning in Pathology*, 149–173.

Campbell, J. B., & Wynne, R. H. (2011). Introduction to remote sensing. Guilford Press.

Congedo, L. (2016). Semi-automatic classification plugin documentation. *Release, 4*(0.1), 29.

Fukatsu, T., & Nanseki, T. (2009). Monitoring system for farming operations with wearable devices utilized sensor networks. *Sensors*, *9*(8), 6171–6184.

JARS (1993). Remote Sensing Note. *Japan Association on Remote Sensing*. Available at http://www.jars1974.net/pdf/rsnote_e.html

Jeremy, J. (2018). Hyperparameter tuning for machine learning models*. Jeremy Jordan*.

Moyroud, N., & Portet, F. (2018). Introduction to QGIS*. QGIS and generic tools,1*, 1-17.

Nguyen, T. T., Hoang, T. D., Pham, M. T., Vu, T. T., Nguyen, T. H., Huynh, Q.-T., & Jo, J. (2020). Monitoring agriculture areas with satellite images and Deep Learning*. Applied Soft Computing, 95*, 106565.

Rogers, J., & Gunn, S. (2005). Identifying feature relevance using a random forest. In International Statistical and Optimization Perspectives Workshop “Subspace, Latent Structure and Feature Selection”. *Springer, Berlin, Heidelberg*, 173-184.

Rotairo, L., Durante, A. C., Lapitan, P., & Rao, L. N. (2019). Use of remote sensing to estimate paddy area and production: a handbook. *Asian Development Bank*.

Schowengerdt, R. A. (2006). Remote sensing: models and methods for image processing. *Elsevier*.

Shippert, P. (2017). Digital number, radiance, and reflectance. *L3Harris Geospatial*.