Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

SEMARAK ILMU
Journal of
ADVANCED RESEARCH IN APPLIED SCIENCES AND ENGINEERING TECHNOLOGY

Block-scale Oil Palm Yield Prediction Using Machine Learning Approaches Based on Landsat and MODIS Satellite Data

Yuhao Ang [1], Helmi Zulhaidi Mohd Shafri [1,*], Yang Ping Lee[2], Shahrul Azman Bakar[2], Haryati Abidin[2], Shaiful Jahari Hashim[3], Mohd Na'aim Samad[3], Nik Norasma Che'ya[4], Mohd Roshdi Hassan[5], Hwee San Lim[6], Rosni Abdullah[7], Yusri Yusup[8], Syahidah Akmal Muhammad[8], Teh Sin Yin[9], Mohamed Barakat A. Gibril[10]

[1] Department of Civil Engineering and Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), 43400 Serdang, Selangor, Malaysia
[2] Geoinformatics Unit, FGV R&D Sdn Bhd, FGV Innovation Centre, PT23417, Lengkuk Teknologi, 71760 Bandar Enstek, Negeri Sembilan, Malaysia
[3] Department of Computer and Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[4] Department of Agriculture Technology, Faculty of Agriculture, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[5] Department of Mechanical and Manufacturing Engineering, Faculty of Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[6] School of Physics, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, Malaysia
[7] School of Computer Sciences, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, Malaysia
[8] School of Industrial Technology, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, Malaysia
[9] School of Management, Universiti Sains Malaysia (USM), 11800 Gelugor, Penang, Malaysia
[10] GIS & Remote Sensing Center, Research Institute of Sciences and Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates, Saudi Arabia

ARTICLE INFO

Article history:
Received 22 June 2023
Received in revised form 14 February 2024
Accepted 11 April 2024
Available online 12 May 2024

ABSTRACT

Due to environmental threats and weather uncertainty concerns, oil palm yield prediction is crucial for sustaining crop production. This can be achieved through machine learning and utilising remotely sensed data to predict crop yield. However, the comparative studies on remotely sensed data in adopting the machine learning models are still limited due to the data accessibility. Therefore, we compare and evaluate the prediction accuracy between different satellites, namely MODIS and Landsat-7, using machine learning algorithms and the topology of deep neural networks. Random forest and stacking outperformed linear regression, ridge regression, and lasso regression for both Landsat-7 NDVI (R^2= 0.78–0.80; RMSE=1.00- 1.26 tonnes per hectare; MAE=0.77-0.79 tonnes per hectares; MAPE=0.03-0.04 tonnes per hectare) and MODIS NDVI (R^2= 0.60–0.65 tonnes per hectares; RMSE= 2.72–2.81 tonne per hectares; MAE= 1.42-1.55, MAPE= 1.01- 1.02 tonnes per hectares). The Landsat-7 NDVI revealed that neural networks with a deeper network topology (R^2= 0.85; RMSE= 1.42 tonnes per hectare; MAE=0.57 tonnes per hectares; MAPE=0.06 tonnes per hectare) outperformed neural networks with a baseline and broader network topologies in terms of performance. In contrast, MODIS-NDVI revealed that the neural network with a wider network topology had the highest overall prediction accuracy and the lowest prediction error (R^2= 0.75; RMSE= 2.81 tonnes per hectare; MAE=2.27 tonnes per tonnes; MAPE= 0.13). Because

[*] *Corresponding author.*
E-mail address: helmi@upm.edu.my

https://doi.org/10.37934/araset.45.1.90107

of its higher spatial resolution in comparison to MODIS, landsat-7 NDVI used in neural networks with a deep network topology provided the best model performance. Although the use of NDVI as a single input factor may cause uncertainty in some extents, it is an efficient and reliable method for improving yield estimation with the use of medium-resolution satellites, which has important implications for early warning towards the reduction in yield production.

Keywords:

Landsat-7; MODIS; NDVI; machine learning; deep neural network

1. Introduction

Early estimation of oil palm yield is essential for responding quickly to plantation issues. A hectare of oil palm cultivation normally generates 3.3 tonnes of oil, which is much higher than the output of vegetable oils such as soybean, which yields about 0.4 t/ha [1]. Current practices include using ground survey as a yield record. This method was inconsistent and was prone to statistical error and bias [2]. Furthermore, numerous current studies have proved that the climate and agronomic aspects are the main elements in predicting oil palm yield [3-5]. Only a limited number of studies have focused on integrating remote sensing feature and machine learning [6,7]. Therefore, the estimation of oil palm yield with the integration of remote sensing feature and machine learning can be an alternative to solve the complexity of factors that affect the decline in oil palm yield.

Over the last several decades, remote sensing such as satellite images has been used to assess crop growth and health [8]. The crop status can be used as a benchmark for early prediction in order to provide recent data prior to the crop harvesting period. The application of remote sensing for yield monitoring is formulated from its close connection to the canopy leaf area index (LAI) and the fraction of absorbed photosynthetically active radiation (fAPAR) [9]. The relation between net primary production and absorbed photosynthetically active radiation (APAR) is linearly proportional [10,11]. Hence, vegetation indices have the potential to be utilised as an indirect indicator of primary crop productivity [12]. Several past studies have shown that normalised difference vegetation index (NDVI) is a broadly used spectral transformation technique in visible and near-infrared (NIR) regions of the electromagnetic spectrum and is suitable for estimation of crop yield.

Recently, machine learning applications can be used for predicting the crop yield when dealing with large data volume [13]. Nevertheless, the oil palm sector is still underutilising machine learning and deep learning applications that use analytics with high adoption algorithms, input data, features, and model evaluation criteria [14]. Aghighi *et al.,* [15] analysed several machine learning techniques for predicting silage maize production using NDVI obtained from Landsat-8 satellite imageries. Boosted regression tree (BRT), random forest regression, support vector regression (SVR), and gaussian process regression (GPR) methodologies were used and evaluated. BRT outperformed the other methods in this study, with an R-value of more than 0.87, while random forest regression was the most stable method for predicting maize production. Phan *et al.,* [16] conducted a study to predict tea yield with other variables such as MODIS-NDVI and mean temperature using three machine learning algorithms: established standard linear regression, support vector regression, and random forest regression. The result showed that random forest regression achieved highest prediction accuracy in estimating tea yield. In another study, Ang *et al.,* [7] proposed a walk validation time-series technique based on advanced ML such as RF and modified AdaBoost algorithms to estimate oil palm yields. The result indicated that RF model surpassed AdaBoost model in estimating oil palm yield. Conceivably, multiple layers of computation in deep neural networks may be utilised in deep learning models to explore heterogeneous information (e.g., remote sensing data) in order to solve the complex and non-linear relationships with crop yields [17-19].

Spatial resolution is essential when considering the use of satellites for yield prediction [20]. As yet, there is lack of comparative study between the Landsat-NDVI and Modis-NDVI in evaluating the accuracy of the yield prediction model in oil palm. Existing studies often experimented on the NDVI derived from single types of satellite images for predicting the yield. Evidence shows that the NDVI behaviours can be influenced by the spatial resolution of the satellite based on the different sensors applied [21]. Therefore, the main objectives of this study were to: 1. Investigate the NDVI derived from Landsat-7 and MODIS satellites in predicting the oil palm yields; 2. Evaluate different machine learning algorithms and topology of deep neural and compare its performance for the prediction of oil palm yields.

2. Methodology

2.1 Study Area

Our study area comprises 40 blocks in a research plantation located in Pahang state in Malaysia, covering 17-kilometre square. Average yearly rainfall (2005–2018) ranges from 112.2 mm to 224.2 mm. The present study area is fully planted with oil palm (Figure 1).

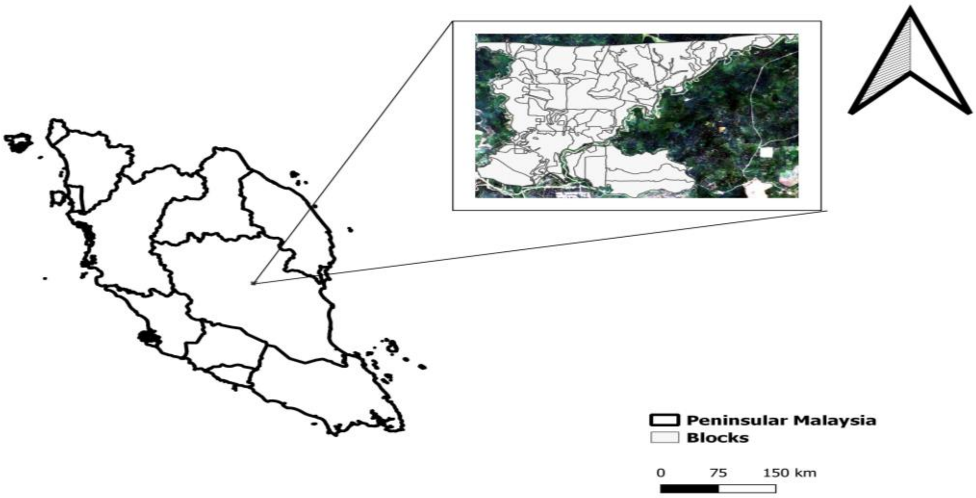


Fig. 1. Location map of the study area

2.2 Yield Data

In this study, we used archive data such as yield information. The plantation administration provided annual yield records. The time span of the analysis was confirmed by the availability of satellite imagery and oil palm yield data. The present study used 40 blocks of yield data from 2005 to 2018 (n= 13 years × 40 blocks = 520 data points) as historical ground information. Averagely, the age of oil palm across blocks ranges from 4 to 28 years. Pre-processing was performed to remove data points with inconsistent yield values, as some blocks are included in the trial plots and may exhibit large variations. The general information on oil palm blocks and basic statistical description of yield from 2005 to 2018 is shown in Table 1.

Table 1
General information on oil palm blocks and basic statistical description of yield

Information	Values
Rainfall	112.2 mm to 224.2 mm
Block sizes in Hectares	1.23 – 145.87
Ages	4 – 28 years old
Count	520
Mean	27.99 tonnes per hectares
Standard deviation	5.21 tonnes per hectares
Max	44.46 tonnes per hectares
Min	18.83 tonnes per hectares
20%	23.26 tonnes per hectares
40%	25.97 tonnes per hectares
60%	28.64 tonnes per hectares
80%	32.52 tonnes per hectares

2.3 Remotely Sensed Data

In order to conduct the analysis for the yield at block level, we used Google Earth Engine to extract an input set of time-series of satellite-based vegetation indices. In this study, a set of time series of Landsat-7 and MODIS imageries in the period of 2005–2018 was selected. Terra moderate resolution spectroradiometer (MODIS) vegetation indices (MOD13Q1 V6) were used every 16 days and 250-meter spatial resolution that provides a vegetation index value at a per pixel basis and processing from atmospherically corrected bi-directional surface reflectance. We used Terra satellite data that is regarded as the continuity index to the existing National Oceanic and Atmospheric Administration-Advanced Very High-Resolution Radiometer (NOAA-AVHRR) derived NDVI. Generated by remote sensing procedures, NDVI is a measure of biomass density on the earth's surface, which can be used to predict crop yields. This index is derived from NIR and red region as shown in the following formula [22].

$$\text{NDVI} = (\text{NIR-RED})/(\text{NIR+RED}) \qquad (1)$$

where NIR = reflectance of Near-infrared; RED is the reflectance of red.

To obtain the precise NDVI values, the pixels in the images were scaled by multiplying by 0.0001. Landsat-7 with a TOA reflectance of level 1 was utilised because the image contains radiometric and geometric corrections for each spectral band. The cloud masking and filling approach was used in pre-processing to remove clouds and fill the region with cloud-free images by employing numerous image combinations at different periods. The reducer mean function was used to calculate the mean NDVI.

2.4 Sampling Strategy

For each block, points were created using Delaunay triangulation, with at least three sampling points connected by neighbouring edge [23]. Given that the multiple possible triangles are defined over neighbouring population elements surrounding a sample point, it is necessary to develop an efficient field protocol that correctly identifies triangles that are part of the same overall triangulation required for the average of each block. Delaunay triangulation was used to determine the natural neighbours of given sampling blocks. Areas with large differences in concentration between natural neighbours may then be targeted for additional sampling blocks, while areas with slight differences may require fewer sampling blocks.

2.5 Machine Learning Workflow

In order to train the model, we divided the data into 70% for training and 30% for validation from 2005 to 2017. In this study, we used Scikit learn and Mixtend for implementing machine learning algorithms for the analysis. Scikit-learn is a library in Python that provides much-unsupervised learning and supervised learning algorithms that can be used for classification purposes. It is efficient to build machine learning models using this library. Mixtend is a Python library consisting of a stacking regressor, which is used to generate the stacking model from multiple based models [24]. One of the advantages is easy implementation with fewer codes.

Traditional linear regression was used in this study as a comparative classifier. Ridge regression and lasso regression which belongs to shrinkage technique were also used [25,26]. Ridge and lasso regressions contain regularisation parameters (alpha), regulate, and penalise. Lasso regression regularisation (L1) uses the magnitude or the vector, that could direct to zero coefficient. Cross-validation approaches and Scikit-learn grid search were used to determine the optimal regulation parameters (Table 2).

Random forest is a bagging method that uses deep trees to fit on bootstrap samples and combine them to produce an output with lower variance. A random subset of the features was selected for weak learners by substituting N examples for each of the features. A weak learner, such as a decision tree, was then fitted for each of these random subspace features and obtained a prediction from each of them before voting to determine the best prediction. Random forest was applied with a few of the parameters. The parameters tunings were aided with randomised grid search. The tunings for random forest, ridge regression, and lasso regression are provided in Table 2.

Table 2
Model hyperparameters and tested values

Information	Values
Ridge regression	Regularisation parameters alpha ranging from 1e-08 to 1e+08
Lasso regression	Regularisation parameters alpha 1e-08 to 1e+08
Random forest	Maximum depth of tree: [10,20,30,40,50,60,70,80,90,100,110] Minimum number of sample leaf required to be at a leaf node: [1,2,4] Minimum number of samples required to split an internal node: [2,5,10] Number of trees in the forest: [200,400,600,800,1000,1200,1400,1600,1800,2000]

Stacking generalisation was applied with fewer based models and combined multiple learning algorithms via meta-learning [27]. Two successive stages of levels needed to be passed through, which are level-0 and level-1. After a series of trial-and-error with multiple selected models, the extreme gradient boosting, linear regression, and lasso regression models with the highest accuracy were chosen as base models in level-0. In level-0, base learners on training data were independently trained and make predictions. In level-1, ridge regression was selected as the meta-learning algorithm that was used to combine the predictions of each base algorithm to produce the best final predictions. A cross validation of these base models was conducted, and then the out-of-fold predictions and the outputs of the base models were used by the meta-regressor to generate a final model for final predictions.

2.6 Topology of Deep Neural Networks

In this study, various topologies of the deep neural networks were investigated to enhance the prediction accuracy, which increases the models' robustness.

First, X_(b,y)^NDVI denotes the NDVI variable NDVI at block b in year y for all NDVI ε {0.3,…,0.7}, b ε {1,…,439}, and y ε {2005,…,2018}. It is important to note that the weight determines the effectiveness of the NDVI based on the years in which the weight is directly proportional to the impact on the network as a whole. The formula is as shown below

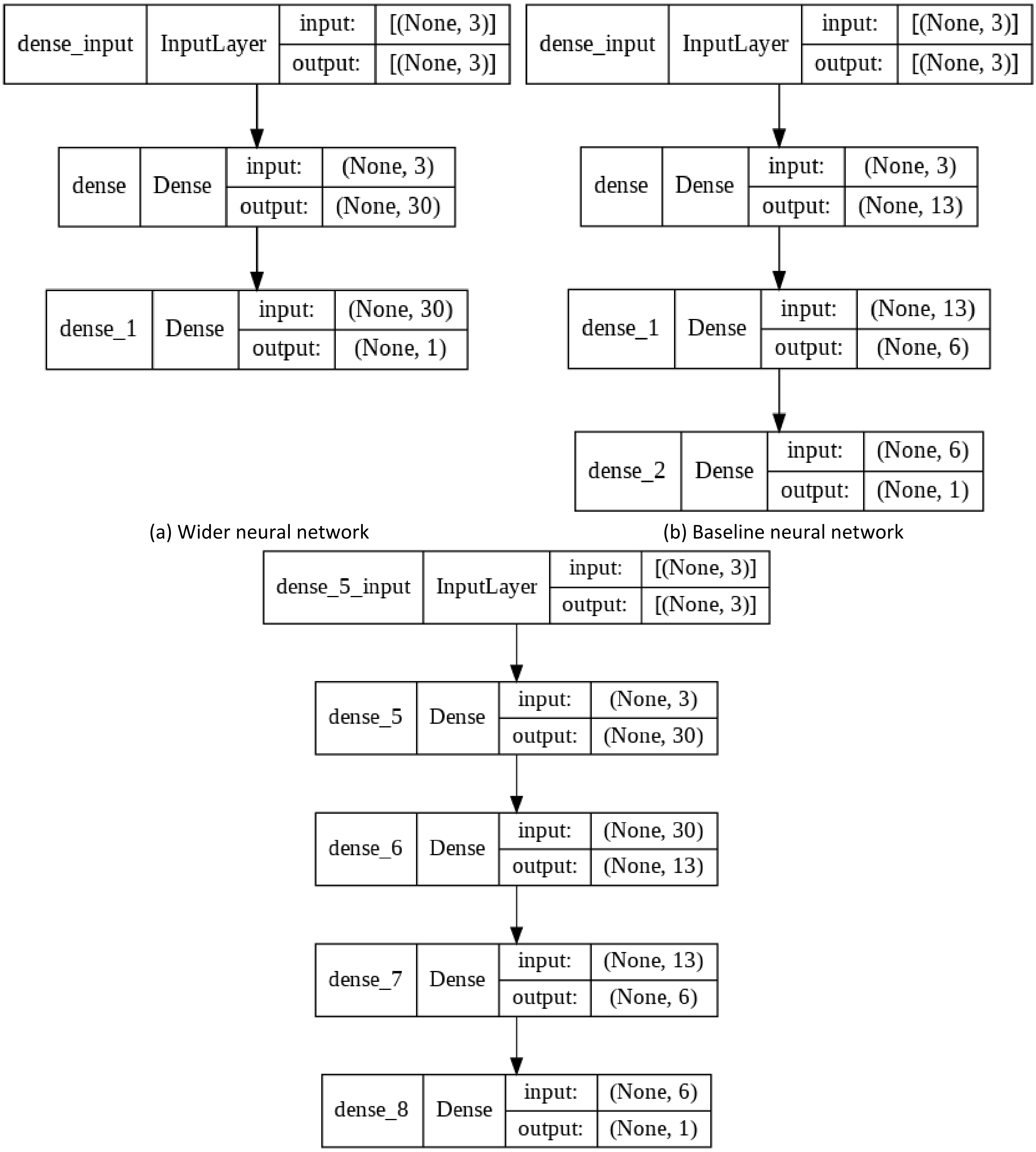
$$f(x)= \sum X_{b,y}^{NDVI} W_i \tag{2}$$

To predict the oil palm yield in 2018, we used historical data from 2005 to 2017 as NDVI variables and we trained three topologies of neural networks, which were used across all blocks. Baseline neural network is a benchmark for this study with one input layers, two hidden layers, and one output layer. In the hidden layers, thirteen and six neurons were used. Activation function transforms the summed weighted input into a smaller value for tiny inputs and applies a threshold to the activation of that output node.

In this study, the rectified linear activation function (RELU) was selected as the default activation function as it has been proven to be effective and easier to train for achieving better accuracy. This function performs mathematical calculations in which neurons are activated based on their output; if the output value declines below zero, the neurons are deactivated from the network. One advantage is that it improves computation efficiency for each parameters update [28,29]. The function of activation function is as shown below

$$f(x)= \begin{cases} x \; if \; x \; positive \\ 0 \; otherwise \end{cases} \tag{3}$$

Neural network with a wider network topology was developed with the number of neurons in the one hidden layer was doubled to thirty. For neural network with a deeper network topology, one input layer, three hidden layers, and one output layer were added to the network. The number of neurons in the hidden layers was set to thirty, thirteen, and six. The learning rate of neural networks was set to 0.01. For Landsat-7 and MODIS satellite imageries, the batch size and epochs were accordingly adjusted in order to optimise the accuracy of the model. The neural network topology is shown in Figure 2.


dense_input | InputLayer | input: [(None, 3)] | output: [(None, 3)]
dense | Dense | input: (None, 3) | output: (None, 30)
dense_1 | Dense | input: (None, 30) | output: (None, 1)
dense_input | InputLayer | input: [(None, 3)] | output: [(None, 3)]
dense | Dense | input: (None, 3) | output: (None, 13)
dense_1 | Dense | input: (None, 13) | output: (None, 6)
dense_2 | Dense | input: (None, 6) | output: (None, 1)
dense_5_input | InputLayer | input: [(None, 3)] | output: [(None, 3)]
dense_5 | Dense | input: (None, 3) | output: (None, 30)
dense_6 | Dense | input: (None, 30) | output: (None, 13)
dense_7 | Dense | input: (None, 13) | output: (None, 6)
dense_8 | Dense | input: (None, 6) | output: (None, 1)

(a) Wider neural network

(b) Baseline neural network

(c) Deeper neural network

Fig. 2. The topology of neural networks used in this study (a) Neural network (wider network topology) with one input layer, one hidden layer with thirty neurons, and one output layer (b) A neural network (baseline network topology) consisting of one input layer, two hidden layers consisting of thirteen and sixteen neurons, and one output layer (c) Neural network (deeper network topology) with one input layer, three hidden layers of thirty, thirteen, and six neurons, and one output layer

2.7 Evaluation of Machine Learning Models

To assess the model performance, we evaluated and predicted the yield for 2018. Four evaluation metrics were computed and assessed from the comparison of predicted and observed yields: R-squared value (R^2), the root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). The errors metrics were computed on yield forecast at block levels. The formulae of the R^2, RMSE, MAPE, and MAE are as follows [30-32]:

R^2=1-1- 〖SS〗 _regression/ 〖SS〗 _total (4)

where 〖SS〗 _regression is known as the sum squared regression error, and 〖SS〗 _total is the sum squared total error.

RMSE √((∑_(i=1)^n▒(X_(i-) y_i)^2)/n) (5)

MAPE= 1/N ∑_t^n▒ 〖|(A_t-P_t)/A_t 〗 | (6)

MAE= (∑_(i=1)^n |y_i-𝓍_i |)/n (7)

We determined the optimal model configuration by comparing the predicted and actual yield values. We used hypothesis testing (a two-sample Z-test with significance level of 5%) to establish the relationship between actual and predicted yield values in machine learning algorithms and deep neural networks for the comparison. Assuming that actual and predicted yield are two distinct data sets for each machine learning algorithm and the topology of deep neural networks, the null hypothesis is that the two data sets are equal in mean.

3. Results

3.1 The Trend of NDVI for Oil Palm

As determined by kernel density estimation, the MODIS NDVI is not normally distributed, whereas the Landsat-7 is normally distributed. The smoothness of the kernel density estimate demonstrates how it estimates for continuous random variables converge more quickly to the actual underlying density. Figure 3 depicts NDVI means, and standard deviation calculated based on NDVI for all blocks (n=38) from 2005 to 2018 using Landsat. The average NDVI value was 0.47. The maximum NDVI value in 2009 was 0.65, while the minimum NDVI value in 2005 was 0.36.

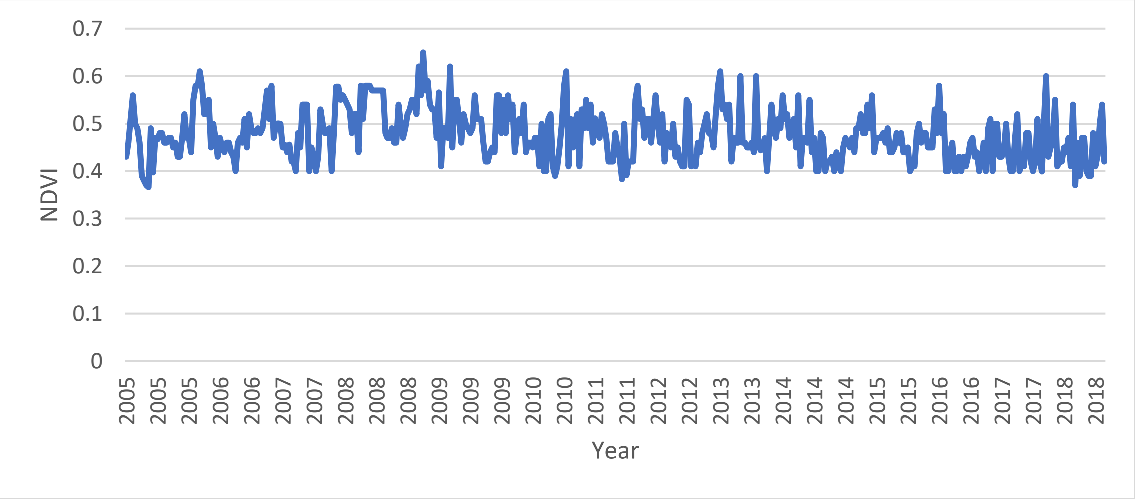


Fig. 3. Time series Landsat-7 NDVI

Figure 4 depicts NDVI means, and standard deviation calculated based on NDVI for all blocks (n=38) from 2005 to 2018 using MODIS. The NDVI value was 0.47 on average. The maximum NDVI value for 2009 was 0.67, while the minimum NDVI value for 2017 was 0.31.

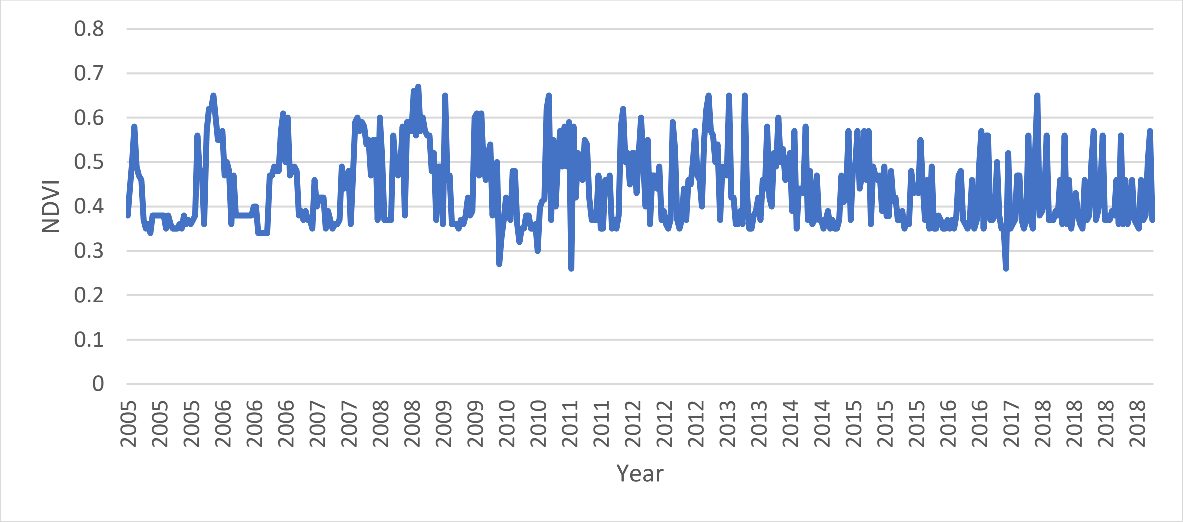


Fig. 4. Time-series MODIS NDVI

3.2 Evaluation of the Machine Learning Algorithms and Neural Network Topology

Table 3 depicts the performance of machine learning models using Landsat data. Machine learning algorithms (linear regression, ridge regression, lasso regression, random forest, and stacking) guarantee promising yield prediction models for oil palm based on highest prediction accuracy. The stacking had the highest overall prediction accuracy with the lowest prediction error, followed by random forest, ridge, and lasso regressions.

Table 3
Performance of the machine learning models using Landsat data

Models	Testing			
	R^2	RMSE (tonnes per hectares)	MAE (tonnes per hectares)	MAPE (tonnes per hectares)
Linear regression	0.50	1.14	0.83	0.03
Ridge regression	0.72	1.07	0.77	0.03
Lasso regression	0.75	1.14	0.83	0.03
Random forest	0.78	1.00	0.79	0.03
Stacking	0.80	1.00	0.77	0.03

In general, machine learning algorithms (Figure 5) produced the highest R-squared values with more than 0.70. This means that the selected machine learning model configuration implied more than 70% of spatial variability for yield prediction in oil palm. Stacking recorded highest R-squared value (R^2= 0.80) and the least prediction errors based on the testing (RMSE=1.00; MAE=0.77 & MAPE=0.03).

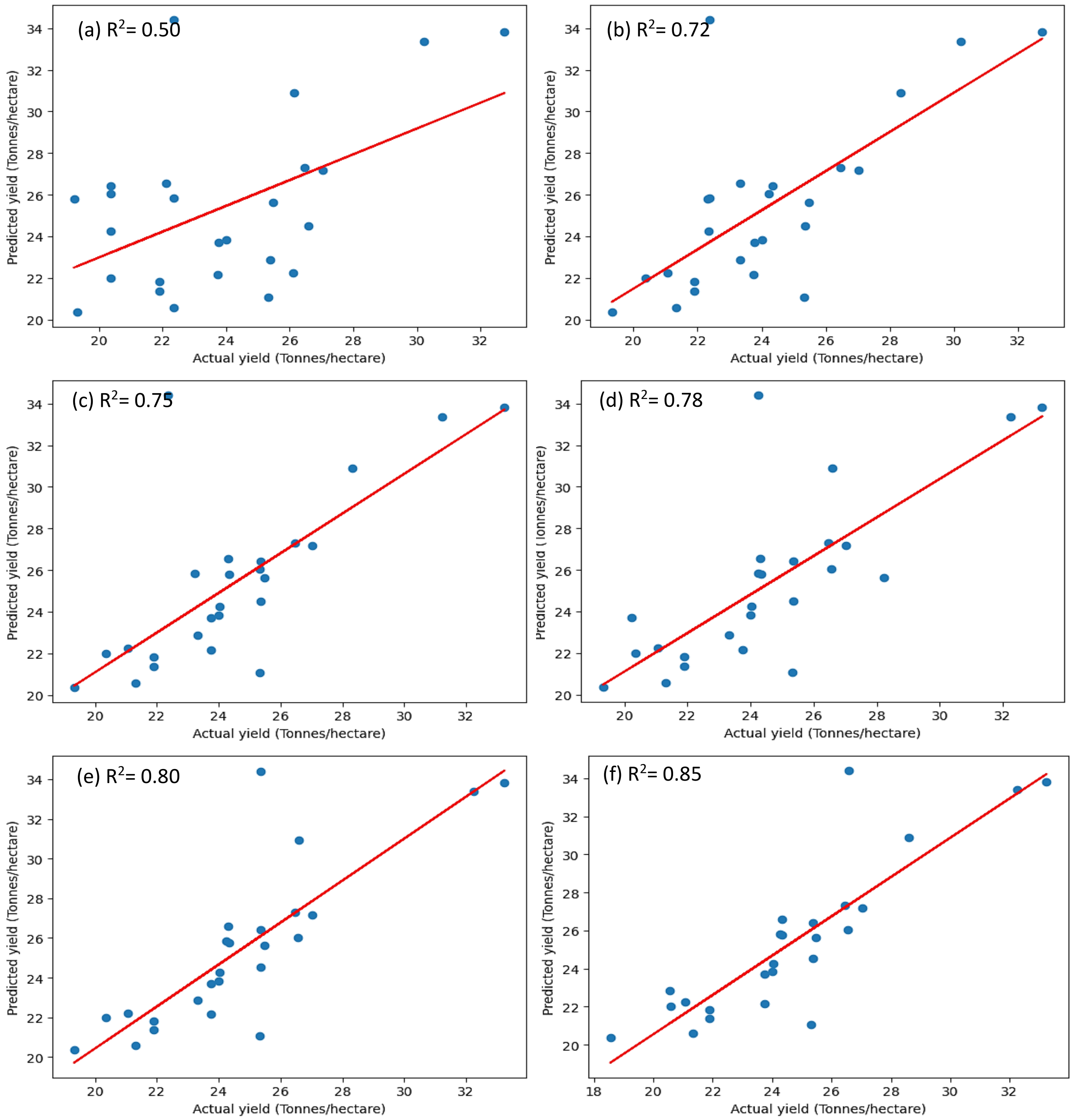


Fig. 5. The machine learning algorithms for developing yield prediction model using Landsat-7 data (a) linear regression (b) lasso regression (c) ridge regression (d) random forest (e) stacking (f) neural network (deeper network topology)

Table 4 depicts the prediction accuracy of machine learning models using MODIS data. Stacking and random forest algorithms achieved the highest R-squared values with around 0.60–0.65, implying that these models had about 60–65% of spatial variability for yield prediction in oil palm. Stacking had the least prediction errors based on the testing (RMSE=2.72; MAE=1.42 & MAPE=0.13). Linear, ridge, and lasso regressions had the highest prediction errors in testing. Stacking algorithms (Figure 6) generated the highest overall prediction accuracy with the lowest prediction error. Random forest achieved the second highest overall prediction accuracy, followed by the ridge, lasso, and linear regressions.

Table 4
Performance of the machine learning models using MODIS data

Models	Testing			
	R^2	RMSE (tonnes per hectares)	MAE (tonnes per hectares)	MAPE (tonnes per hectares)
Linear regression	0.20	4.11	3.32	0.13
Ridge regression	0.32	4.12	3.33	0.13
Lasso regression	0.34	4.11	3.32	0.13
Random forest	0.60	2.81	1.55	0.13
Stacking	0.65	2.72	1.42	0.13

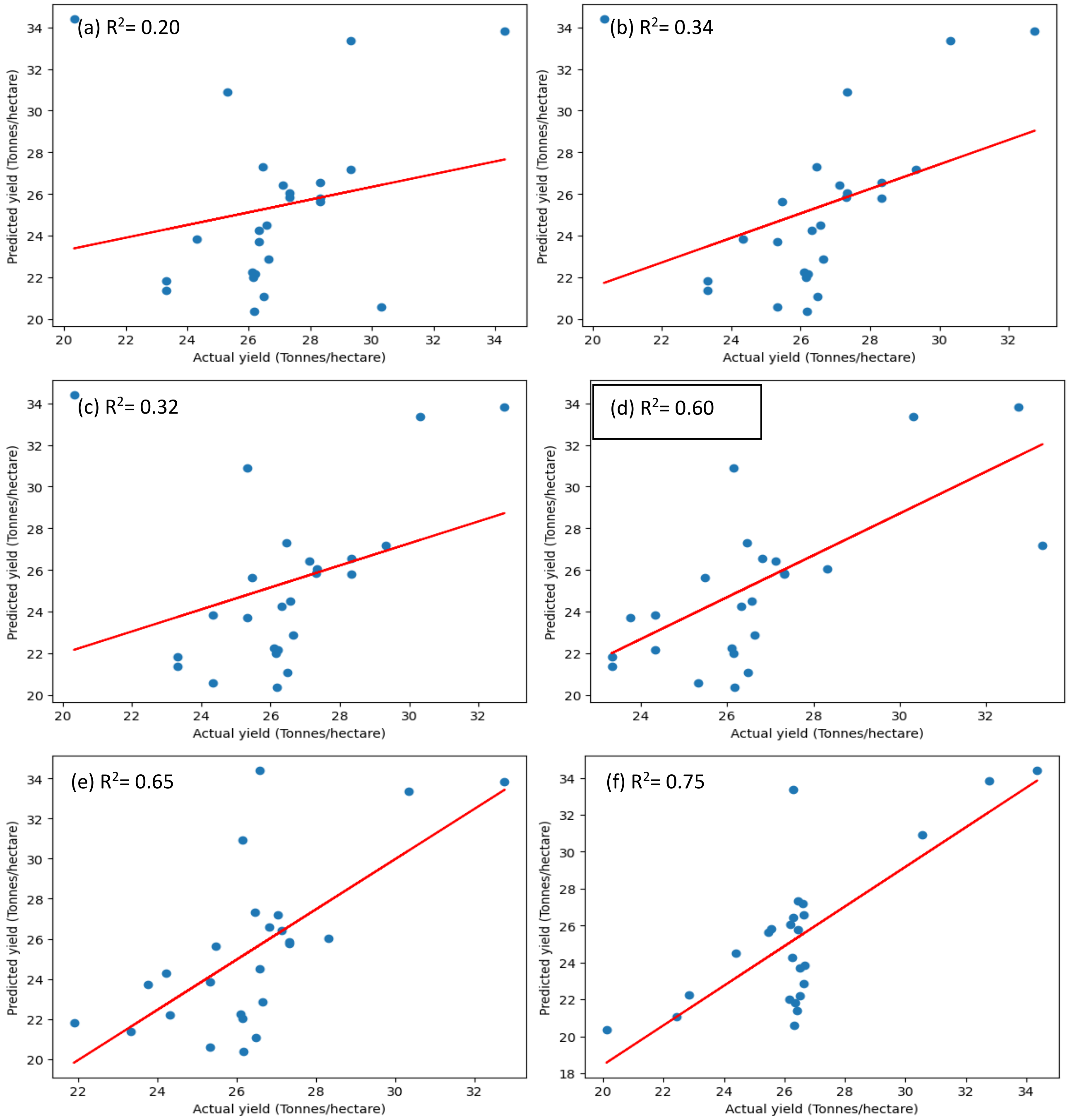


Fig. 6. The machine learning algorithms for developing yield prediction model using MODIS data (a) linear regression (b) lasso regression (c) ridge regression (d) random forest (e) stacking (f) neural network (wider network topology)

Table 5 depicts the evaluation of the deep neural network using Landsat-7 data. It was found that deeper neural networks produced the highest overall prediction accuracy with the lowest error, followed by baseline neural networks and wider neural networks. Neural network with a wider, baseline and deeper network topologies achieved the highest R-squared values with around 0.77-0.85, denoting that these models implied approximately 77-85% of spatial variability for oil palm yield prediction. Neural network with a deeper network topology (Figure 5(f)) yielded best overall prediction accuracy and the lowest prediction errors in testing (RMSE=1.42; MAE=0.57; MAPE=0.06).

Table 5
Performance of the deep neural networks using Landsat-7 data

Types	Number of neurons	Number of hidden layers	Testing			
			R^2	RMSE (tonnes per hectares)	MAE (tonnes per hectares)	MAPE (tonnes per hectares)
Wider neural network	30	1	0.79	1.80	1.52	0.06
Baseline neural network	13, 6	2	0.77	1.56	4.82	0.19
Deeper neural network	30, 13, 6	4	0.85	1.42	0.57	0.06

Table 6 shows the evaluation of the neural network using MODIS data. Neural network with a wider network topology had the highest overall prediction accuracy with the lowest error, followed by neural network with baseline and deeper network topologies. Neural networks with a baseline and wider network topologies had 73% and 75% spatial variability for yield prediction in oil palm, deduced from their high R-squared values of 0.73 and 0.75, respectively. Neural network with a wider network (Figure 6(f)) generated the best overall prediction accuracy and lowest prediction errors in testing (RMSE=2.81; MAE=2.27; MAPE=0.13). Whereas Neural network with a deeper network topology achieved the lowest R-squared value of 0.52, implying about 52% of spatial variability for yield prediction in oil palm.

Table 6
Performance of the deep neural networks using MODIS data

Types	Number of neurons	Number of hidden layers	Testing			
			R^2	RMSE (tonnes per hectares)	MAE (tonnes per hectares)	MAPE (tonnes per hectares)
Wider neural network	30	1	0.75	2.81	2.27	0.13
Baseline neural network	13,6	2	0.73	3.87	3.05	0.13
Deeper neural network	30,13,6	4	0.52	6.80	5.51	0.14

Although the most effective machine learning algorithms produce the most accurate forecasts in estimating oil palm yields, differences between other algorithms are not always statistically significant. For Landsat-NDVI, results showed that random forest and stacking regressions are the best machine learning algorithms. The null hypothesis was accepted, showing no difference since it did not reach the significance level of 5%. In contrast, neural networks with wider and deeper network topologies are acceptable with the least prediction accuracy errors with no difference as it did not reach significance level of 5%. The neural network with the baseline network topology showed a significant difference with a significance level of 5%, so it is not appropriate to use it for oil palm yield prediction.

Random forest and stacking algorithms did not show any significant differences for MODIS NDVI since it did not reach 5% significance levels, which indicates that these models are appropriate for this study. It was found that a neural network with a wider topology showed no differences since the 5% level of significance was not reached.

4. Discussion

Stacking and the random forest algorithms had the best overall prediction accuracy and the lowest prediction error in our study. Previous studies have shown that ensemble learning algorithms were beneficial for crop yield prediction [33,34]. In agreement with Shahhosseini *et al.,* [35], we observed that stacking regression achieved higher accuracy than random forest. In a similar study, Nishant *et al.,* [36] proved that stacking regression achieved the highest prediction accuracy in predicting all kinds of crops that are planted in India. One of the reasons is that combining and averaging multiple base models to produce the final prediction improves accuracy compared to single model alone [37]. Our findings are in line with other studies conducted by Chandra *et al.,* [38] and Wen *et al.,* [39], which show that random forest can be used to predict the yield. Phan *et al.,* [40] reported that a random forest model was better than a support vector machine model at predicting tea yield using MODIS-NDVI with R-squared between 0.67–0.71.

Overall, a deep neural network provides reliable accuracy for predicting yield in oil palm. As we continued the work studied by Khaki and Wang [41], we investigated the topology of the neural network. Hara *et al.,* [32] reported that one hidden layer and four neurons were the best configuration to predict seed yield accurately. Our work extended the studies of Haque *et al.,* [42], which used only three hidden layers for predicting the crop yield. We found that neural networks with wider and deeper network topologies can improve performance depending on the configuration settings, such as the number of neurons and hidden layers. The number of hidden layers and hidden units in a deeper neural network can be affected by the total number of inputs, the complexity of the deep neural network structure, the number of samples used in training, the amount of noise in the sample set, the output units, and the training algorithm [43-45]. For example, the Landsat-7 NDVI showed that neural networks with deeper network topologies outperformed those with wider and baseline network topologies, whereas the MODIS-NDVI showed that neural networks with wider network topologies outperformed those with deeper and baseline topologies. This indicated that increasing the number of layers did not enhance accuracy, but rather greatly increased the training complexity. Deep neural networks, as opposed to standard linear regression, can predict NDVI behaviour changes in relation to yield since MODIS-NDVI contains generalisation issue due to lower spatial resolution for the prediction of yield at block level, which has caused the results more complicated. Upon investigating the influence of the neural network topology on prediction accuracy in improving the model, Naitzat *et al.,* [46] suggested that the increase or decrease in the width and depth, respectively, on topology change will increase the training patterns and thus affect the accuracy. Our result was in agreement with Aghighi *et al.,* [15], who demonstrated that the ML approaches outperformed traditional regression methods. One of the reasons is that ML approaches outperformed traditional regression methods owing to their capacity to cope with high-dimensional data of complicated distributions as well as the inconsistency of NDVI time series.

In this study, we identified Landsat-7 was contributing a higher accuracy with minimal prediction error in yield prediction compared to the MODIS. This is due to the higher resolution pixels which contained a more significant fraction of the agricultural target [47]. This study revealed that a greater spatial resolution at the block level, rather than a higher revisit frequency, accredits more accurate yield prediction. Although many studies have shown promising results for yield prediction using MODIS [48,49] at regional and country levels, it is not feasible to predict the yields at the block level. Our study was in agreement with Jurečka *et al.,* [50]. In that comparative study on spatial yield variability with MODIS and Landsat product, it was shown that the correlation of MODIS-NDVI and yield is lower with r=0.1 and inconsistent in the prediction, as compared to Landsat-NDVI with the r> 0.751 [50].

Furthermore, Van *et al.,* [51] proved that MODIS-NDVI data were influenced by atmospheric water vapour, and the impact was significant. The result of our study nearly agrees with other studies that compared the accuracy of the spatial resolutions of satellites. This could be due to noises such as atmospheric effects and may not accurately reflect the actual scenario for yield prediction at the block-scale. Durgun *et al.,* [52] studied spatial resolution on wheat yield prediction using PROBA-V satellite with 100m, 300m, and 1 km resolutions. The results revealed that PROBA-V satellite with 100m pixel resolution provided more accurate estimates of wheat yield estimation with adjusted R^2=0.74, RMSE=0.6 t/ha, and MAE=0.46 t/ha. The particularity of this present study compared to the aforementioned studies is that this study investigated the prediction accuracy of the Landsat-7 dataset and MODIS datasets using adopted machine learning algorithms. The reason for poor performance using MODIS data is generalisation due to lower spatial resolution than Landsat data.

Our methodology was limited to one variable, which is NDVI. Added relevant variables are more likely to increase prediction accuracy. For instance, weather variables such as temperature, solar radiation, rainfall, and precipitation can increase the prediction accuracy when these variables are combined with NDVI [16,53,54]. Other vegetation indices help to minimise soil and atmospheric disturbances and avoid NDVI saturation at dense canopy cover [55]. Therefore, a variety of vegetation indices will be considered in future studies. Future research may focus on investigating these issues using higher resolution satellite images. It is most likely that the inaccuracies of the model are due to a significant noise signal on NDVI signals generated by cloud cover.

Increasing temporal resolution may also improve crop growth and the accuracy in yield estimation [56]. However, some open-source data have a 10-year archive limit, but access to historical observations is essential for constructing a sufficiently large training set for yield estimation. In this study, thirteen years of collected data resulted in 439 data points based on block management that meet machine learning standards. In future studies, deep learning algorithms will be adopted when handling more data points to ensure the data veracity when maintaining the model accuracy.

5. Conclusion

Using MODIS and Landsat-7 satellites, we developed an effective machine learning framework for forecasting oil palm yields. Our research may be further automated to select the optimum model for predicting oil palm yields. A rigorous testing procedure is utilised to provide blocks that are suitable for sustained production of oil palm. Overall, NDVI derived from Landsat-7 can be used for oil palm yield prediction in the block level using machine learning algorithms and deep neural network topology. The best model performance was obtained by using landsat-7 NDVI with a deeper network topology as it has a higher spatial resolution than MODIS. However, MODIS-NDVI is not recommended for the block level due to the highest prediction errors and the lesser spatial resolution. Therefore, the selection of the satellites should be considered based on the spatial and temporal resolutions, which can accurately predict the yields over the size of the study area. Due to the inconsistency of NDVI time series by satellites, there is a shift toward advanced machine learning, including deep learning. Our result also proved that different neural network topology may affect prediction accuracy. For future studies, it is necessary to collect and incorporate additional yield data and important factors spanning more than 18 years to make more precise predictions.

Acknowledgement

The researchers would like to thank the Malaysian Ministry of Higher Education (MOHE) for funding and resources for this study via the Long-Term Research Grant Scheme (LRGS)-Malaysia Research University Network (MRUN). This project is being carried out jointly with the research program entitled 'A Big Data Analytics Platform for Optimizing Oil Palm Yield Via Breeding by Design (Grant No: 203.PKOMP.6770007)' and the project is entitled "Geoinformatics Data for Palm Oil Yield Prediction Using Machine Learning" (Vote No: 6300268-10801). The authors would also like to thank the team at FGV R&D Sdn. Bhd. for their assistance in assisting with this research.

References

[1] WWF. "8 Things to Know About Palm Oil." *WWF-UK* (2020).

[2] Shamshiri, Redmond Ramin, Ibrahim A. Hameed, Siva K. Balasundram, Desa Ahmad, Cornelia Weltzien, and Muhammad Yamin. "Fundamental research on unmanned aerial vehicles to support precision agriculture in oil palm plantations." *Agricultural Robots-Fundamentals and Application* (2018): 91-116.

[3] Kartika, Nadia Dwi, I. Wayan Astika, and Edi Santosa. "Oil palm yield forecasting based on weather variables using artificial neural network." *Indonesian Journal of Electrical Engineering and Computer Science* 3, no. 3 (2016): 626-633. https://doi.org/10.11591/ijeecs.v3.i3.pp626-633

[4] Oettli, Pascal, Swadhin K. Behera, and Toshio Yamagata. "Climate based predictability of oil palm tree yield in Malaysia." *Scientific reports* 8, no. 1 (2018): 2271.https://doi.org/10.1038/s41598-018-20298-0

[5] Siang, Cheah See, Christopher Teh Boon Sung, Mohd Razi Ismail, and Mohd Rafii Yusop. "Modelling hourly AIR temperature, relative humidity and solar irradiance over several major oil palm growing areas in Malaysia." *J. Oil Palm Res* 32 (2020): 34-49.https://doi.org/10.21894/jopr.2020.0010

[6] Diana, Shinta Rahma, Syaiful Muflichin Purnama, Gusti Dharma, Agil Sutrisnanto, Intan Perwitasari, and Farida Farida. "Estimation the amount of oil palm production using Artificial Neural Network and NDVI SPOT-6 Imagery." *International Journal of Innovative Science and Research Technology* 4, no. 11 (2019): 548–554.

[7] Ang, Yuhao, Helmi Zulhaidi Mohd Shafri, Yang Ping Lee, Haryati Abidin, Shahrul Azman Bakar, Shaiful Jahari Hashim, Nik Norasma Che'Ya, Mohd Roshdi Hassan, Hwee San Lim, and Rosni Abdullah. "A novel ensemble machine learning and time series approach for oil palm yield prediction using Landsat time series imagery based on NDVI." *Geocarto International* 37, no. 25 (2022): 9865-9896. https://doi.org/10.1080/10106049.2022.2025920

[8] Sishodia, Rajendra P., Ram L. Ray, and Sudhir K. Singh. "Applications of remote sensing in precision agriculture: A review." *Remote sensing* 12, no. 19 (2020): 3136. https://doi.org/10.3390/rs12193136

[9] Řezník, Tomáš, Tomáš Pavelka, Lukáš Herman, Vojtěch Lukas, Petr Širůček, Šimon Leitgeb, and Filip Leitner. "Prediction of yield productivity zones from Landsat 8 and Sentinel-2A/B and their evaluation using farm machinery measurements." *Remote Sensing* 12, no. 12 (2020): 1917. https://doi.org/10.3390/rs12121917

[10] Bolton, Douglas K., and Mark A. Friedl. "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics." *Agricultural and Forest Meteorology* 173 (2013): 74-84. https://doi.org/10.1016/j.agrformet.2013.01.007

[11] Liang, S., and J. Wang. "Chapter 11—Fraction of absorbed photosynthetically active radiation." *Advanced Remote Sensing, 2nd ed.; Academic Press: Cambridge, MA, USA* (2020): 447-476. https://doi.org/10.1016/b978-0-12-815826-5.00011-8

[12] Xue, Jinru, and Baofeng Su. "Significant remote sensing vegetation indices: A review of developments and applications." *Journal of sensors* 2017 (2017). https://doi.org/10.1155/2017/1353691

[13] Liakos, Konstantinos G., Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. "Machine learning in agriculture: A review." *Sensors* 18, no. 8 (2018): 2674. https://doi.org/10.3390/s18082674

[14] Khan, Nuzhat, Mohamad Anuar Kamaruddin, Usman Ullah Sheikh, Yusri Yusup, and Muhammad Paend Bakht. "Oil palm and machine learning: Reviewing one decade of ideas, innovations, applications, and gaps." *Agriculture* 11, no. 9 (2021): 832. https://doi.org/10.3390/agriculture11090832

[15] Aghighi, Hossein, Mohsen Azadbakht, Davoud Ashourloo, Hamid Salehi Shahrabi, and Soheil Radiom. "Machine learning regression techniques for the silage maize yield prediction using Time-Series images of Landsat 8 OLI." *IEEE*

Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, no. 12 (2018): 4563–4577. https://doi.org/10.1109/JSTARS.2018.2823361

[16] Phan, Phamchimai, Nengcheng Chen, Lei Xu, and Zeqiang Chen. "Using multi-temporal MODIS NDVI data to monitor tea status and forecast yield: A case study at Tanuyen, Laichau, Vietnam." *Remote Sensing* 12, no. 11 (2020): 1814. https://doi.org/10.3390/rs12111814

[17] Dharani, M. K., R. Thamilselvan, P. Natesan, P. C. D. Kalaivaani, and S. Santhoshkumar. "Review on crop prediction using deep learning techniques." In *Journal of Physics: Conference Series*, vol. 1767, no. 1, p. 012026. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1767/1/012026

[18] Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." *Computers and Electronics in Agriculture* 177 (2020): 105709. https://doi.org/10.1016/j.compag.2020.105709

[19] Kamilaris, Andreas, and Francesc X. Prenafeta-Boldú. "Deep learning in agriculture: A survey." *Computers and electronics in agriculture* 147 (2018): 70-90. https://doi.org/10.1016/j.compag.2018.02.016

[20] Rembold, Felix, Clement Atzberger, Igor Savin, and Oscar Rojas. "Using low resolution satellite imagery for yield prediction and yield anomaly detection." *Remote Sensing* 5, no. 4 (2013): 1704-1733. https://doi.org/10.3390/rs5041704

[21] Huang, Sha, Lina Tang, Joseph P. Hupy, Yang Wang, and Guofan Shao. "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing." *Journal of Forestry Research* 32, no. 1 (2021): 1–6. https://doi.org/10.1007/s11676-020-01155-1

[22] Rouse Jr, John W., R. Hect Haas, D. W. Deering, J. A. Schell, and James C. Harlan. *Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation*. No. E75-10354. 1974.

[23] Wang, Wei. "Sampling and predicting geographic areas using participatory sensing." (2015).

[24] Raschka, Sebastian. "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack." *Journal of open source software* 3, no. 24 (2018): 638. https://doi.org/10.21105/joss.00638

[25] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58, no. 1 (1996): 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[26] McDonald, G.C. "Ridge regression." *Wiley Interdisciplinary Reviews: Computational Statistics* 1, no. 1 (2009): 93-100. https://doi.org/10.1002/wics.14

[27] Breiman, Leo. "Stacked regressions." *Machine Learning* 24, no. 1 (1996): 49-64. https://doi.org/10.1023/A:1018046112532

[28] Eckle, Konstantin, and Johannes Schmidt-Hieber. "A comparison of deep networks with ReLU activation function and linear spline-type methods." *Neural Networks* 110 (2019): 232-242. https://doi.org/10.1016/j.neunet.2018.11.005

[29] Schmidt-Hieber, Johannes. "Nonparametric regression using deep neural networks with ReLU activation function." *The Annals of Statistics* 48, n0. 4 (2020): 1875-1897. https://doi.org/10.1214/19-AOS1875

[30] Chen, Jeng-Fung, Quang Hung Do, Thi Van Anh Nguyen, and Thi Thanh Hang Doan. "Forecasting monthly electricity demands by wavelet neuro-fuzzy system optimized by heuristic algorithms." *Information* 9, no. 3 (2018): 51. https://doi.org/10.3390/info9030051

[31] Schwalbert, Raí A., Telmo Amado, Geomar Corassa, Luan Pierre Pott, PV Vara Prasad, and Ignacio A. Ciampitti. "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil." *Agricultural and Forest Meteorology* 284 (2020): 107886. https://doi.org/10.1016/j.agrformet.2019.107886

[32] Hara, Patryk, Magdalena Piekutowska, and Gniewko Niedbała. "Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data." *Land* 10, no. 6 (2021): 609. https://doi.org/10.3390/land10060609

[33] Li, Changchun, Yilin Wang, Chunyan Ma, Weinan Chen, Yacong Li, Jingbo Li, Fan Ding, and Zhen Xiao. "Improvement of Wheat Grain Yield Prediction Model Performance Based on Stacking Technique." *Applied Sciences* 11, no. 24 (2021): 12164. https://doi.org/10.3390/app112412164

[34] Arumugam, Ponraj, Abel Chemura, Bernhard Schauberger, and Christoph Gornott. "Remote sensing based yield estimation of rice (Oryza sativa l.) using gradient boosted regression in India." *Remote Sensing* 13, no. 12 (2021): 2379. https://doi.org/10.3390/rs13122379

[35] Shahhosseini, Mohsen, Guiping Hu, and Sotirios V. Archontoulis. "Forecasting corn yield with machine learning ensembles." *Frontiers in Plant Science* 11 (2020): 527890. https://doi.org/10.3389/fpls.2020.01120

[36] Nishant, Potnuru Sai, Pinapa Sai Venkat, Bollu Lakshmi Avinash, and B. Jabber. "Crop yield prediction based on Indian agriculture using machine learning." In *2020 international conference for emerging technology (INCET)*, pp. 1-4. IEEE, 2020. https://doi.org/10.1109/INCET49848.2020.9154036

[37] Shahhosseini, Mohsen, Guiping Hu, Isaiah Huber, and Sotirios V. Archontoulis. "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt." *Scientific reports* 11, no. 1 (2021): 1606. https://doi.org/10.1038/s41598-020-80820-1

[38] Chandra, Aditi, Pabitra Mitra, S. K. Dubey, and S. S. Ray. "Machine learning approach for kharif rice yield prediction integrating multi-temporal vegetation indices and weather and non-weather variables." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2019): 187-194. https://doi.org/10.5194/isprs-archives-XLII-3-W6-187-2019

[39] Wen, Guoqi, Bao-Luo Ma, Anne Vanasse, Claude D. Caldwell, Hugh J. Earl, and Donald L. Smith. "Machine learning-based canola yield prediction for site-specific nitrogen recommendations." *Nutrient Cycling in Agroecosystems* 121, no. 2 (2021): 241-256. https://doi.org/10.1007/s10705-021-10170-5

[40] Phan, Phamchimai, Nengcheng Chen, Lei Xu, Duy Minh Dao, and Dinhkha Dang. "NDVI variation and yield prediction in growing season: a case study with tea in Tanuyen Vietnam." *Atmosphere* 12, no. 8 (2021): 962. https://doi.org/10.3390/atmos12080962

[41] Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." *Frontiers in plant science* 10 (2019): 452963. https://doi.org/10.3389/fpls.2019.00621

[42] Haque, Fatin Farhan, Ahmed Abdelgawad, Venkata Prasanth Yanambaka, and Kumar Yelamarthi. "Crop yield prediction using deep neural network." In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pp. 1-4. IEEE, 2020. https://doi.org/10.1109/WF-IoT48130.2020.9221298

[43] Movagharnejad, Kamyar, and Maryam Nikzad. "Modeling of tomato drying using artificial neural network." *Computers and electronics in agriculture* 59, no. 1-2 (2007): 78-85. https://doi.org/10.1016/j.compag.2007.05.003

[44] Panda, Sudhanshu Sekhar, Daniel P. Ames, and Suranjan Panigrahi. "Application of vegetation indices for agricultural crop yield prediction using neural network techniques." *Remote sensing* 2, no. 3 (2010): 673-696. https://doi.org/10.3390/rs2030673

[45] Zhang, Hao, Hao Hu, Xiao-bin Zhang, Lian-feng Zhu, Ke-feng Zheng, Qian-yu Jin, and Fu-ping Zeng. "Estimation of rice neck blasts severity using spectral reflectance based on BP-neural network." *Acta Physiologiae Plantarum* 33 (2011): 2461-2466. https://doi.org/10.1007/s11738-011-0790-0

[46] Naitzat, Gregory, Andrey Zhitnikov, and Lek-Heng Lim. "Topology of deep neural networks." *Journal of Machine Learning Research* 21, no. 184 (2020): 1-40.

[47] Chen, Pei-Yu, Gunar Fedosejevs, Mario Tiscareño-LóPez, and Jeffrey G. Arnold. "Assessment of MODIS-EVI, MODIS-NDVI and VEGETATION-NDVI composite data using agricultural measurements: An example at corn fields in western Mexico." *Environmental monitoring and assessment* 119 (2006): 69-82. https://doi.org/10.1007/s10661-005-9006-7

[48] Panek, Ewa, and Dariusz Gozdowski. "Analysis of relationship between cereal yield and NDVI for selected regions of Central Europe based on MODIS satellite data." *Remote Sensing Applications: Society and Environment* 17 (2020): 100286. https://doi.org/10.1016/j.rsase.2019.100286

[49] Panek, Ewa, and Dariusz Gozdowski. "Relationship between MODIS derived NDVI and yield of cereals for selected European countries." *Agronomy* 11, no. 2 (2021): 340. https://doi.org/10.3390/agronomy11020340

[50] Jurečka, František, Vojtěch Lukas, Petr Hlavinka, Daniela Semerádová, Zdeněk Žalud, and Miroslav Trnka. "Estimating crop yields at the field level using Landsat and MODIS products." *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* 66, no. 5 (2018). https://doi.org/10.11118/actaun201866051141

[51] Van Leeuwen, Willem JD, Barron J. Orr, Stuart E. Marsh, and Stefanie M. Herrmann. "Multi-sensor NDVI data continuity: Uncertainties and implications for vegetation monitoring applications." *Remote sensing of environment* 100, no. 1 (2006): 67-81. https://doi.org/10.1016/j.rse.2005.10.002

[52] Durgun, Yetkin Özüm, Anne Gobin, Grégory Duveiller, and Bernard Tychon. "A study on trade-offs between spatial resolution and temporal sampling density for wheat yield estimation using both thermal and calendar time." *International Journal of Applied Earth Observation and Geoinformation* 86 (2020): 101988. https://doi.org/10.1016/j.jag.2019.101988

[53] Meroni, Michele, François Waldner, Lorenzo Seguini, Hervé Kerdiles, and Felix Rembold. "Yield forecasting with machine learning and small data: What gains for grains?." *Agricultural and Forest Meteorology* 308 (2021): 108555. https://doi.org/10.1016/j.agrformet.2021.108555

[54] Evans, Fiona H., and Jianxiu Shen. "Long-term hindcasts of wheat yield in fields using remotely sensed phenology, climate data and machine learning." *Remote Sensing* 13, no. 13 (2021): 2435. https://doi.org/10.3390/rs13132435

[55] Peralta, Nahuel R., Yared Assefa, Juan Du, Charles J. Barden, and Ignacio A. Ciampitti. "Mid-season high-resolution satellite imagery for forecasting site-specific corn yield." *Remote Sensing* 8, no. 10 (2016): 848. https://doi.org/10.3390/rs8100848

[56] Waldner, François, Heidi Horan, Yang Chen, and Zvi Hochman. "High temporal resolution of leaf area data improves empirical estimation of grain yield." *Scientific reports* 9, no. 1 (2019): 15714. https://doi.org/10.1038/s41598-019-51715-7