Journal of Oil Palm Research
DOI: https://doi.org/10.21894/jopr.2025.0044

PREDICTION OF OIL PALM YIELD USING MACHINE LEARNING: COMPARISON OF LINEAR AND NON-LINEAR ALGORITHMS WITH MULTIVARIATE TIME SERIES DATA

NUZHAT KHAN[1,2]; MOHAMAD ANUAR KAMARUDDIN[1]*; USMAN ULLAH SHEIKH[2]; AB AL-HADI BIN AB RAHMAN[2] and MUHAMMAD PAEND BAKHT[3]*

ABSTRACT

Oil palm yield prediction plays a vital role in supporting sustainable agricultural practices and guiding strategic decisions in the palm oil industry. With the increasing availability of historical and weather-related data, machine learning has become a promising approach for forecasting crop yields. This study evaluates the performance of both linear and non-linear machine learning models using historical agrometeorological data from 1986-2020 collected in Pahang, Malaysia. Specifically, we compare Linear Regression with three non-linear, tree-based models: Extra Trees, Random Forest and Gradient Boosting. The results show that the Extra Trees outperformed all other models explaining 88% of the variance (R^2) in validation data with the lowest prediction error. Random Forest and Gradient Boosting also demonstrated strong performance with R^2 values of 79% and 78%, respectively. In contrast, Linear Regression achieved an R^2 of only 41%, indicating a limited ability to capture the non-linear relationships inherent in weather and environmental variables. This underperformance highlights the structural limitations of linear models when applied to complex agricultural datasets. Although non-linear models are computationally more demanding, their superior capacity to model complex, non-linear patterns makes them more suitable for real-world agricultural applications. These findings emphasise the value of tree-based machine learning models particularly Extra Trees in delivering reliable and accurate yield predictions, which are essential for sustainable oil palm plantation management.

Keywords: machine learning, oil palm, yield prediction.

Received: 18 October 2024; **Accepted:** 26 June 2025; **Published online:** 20 August 2025.

INTRODUCTION

Industrial Revolution 4.0 (IR4.0) in agriculture refers to the integration of advanced technologies such as big data and artificial intelligence (AI) to optimise farming operations (An-Vo *et al*., 2021). These technologies enable the collection and analysis of data on weather patterns, soil conditions, crop yields, and pest populations to facilitate smarter decision-making and more efficient resource management (Khan *et al*., 2022b). Among these advancements, AI-powered predictive analytics play a crucial role in forecasting crop yields for future planning. They also help farmers to improve food security by adapting to climate variability (Javaid *et al*., 2023).

Crop yield prediction is a complex task influenced by various factors, including genetics, soil conditions, weather and management practices (Khaki *et al*., 2020). Traditional methods, such as field surveys and crop growth models, have limitations

[1] School of Industrial Technology, Universiti Sains Malaysia, 11800 Gelugor, Pulau Pinang, Malaysia.

[2] School of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.

[3] Agricultural Economics and Farm Surveys Department, Teagasc, H65 R718 Galway, Ireland.

* *Corresponding author e-mail: anuarkamaruddin@usm.my, MuhammadPaend.bakht@teagasc.ie*

in capturing the dynamic nature of agricultural environments. Advanced machine learning that integrates traditional methods with data-driven modelling presents a compelling approach to improving forecast accuracy (Gebresenbet *et al*., 2023). The access to agrometeorological data, enhanced computational power, and recent advances in machine learning provide valuable insights into agricultural systems. Many studies have used machine learning to predict yields of several crops, but it is still unclear which data types, preprocessing methods and machine learning models are best suited for accurate crop yield prediction. Moreover, many existing studies are crop or region-specific, limiting the generalisability of their approaches (Bharadiya *et al*., 2023). This limitation necessitates the need to develop, examine and compare different machine learning techniques for yield prediction of individual crops to find appropriate, reliable and robust methods (Rashid *et al*., 2021).

Machine learning models for yield prediction can be broadly categorised into linear and non-linear (Mendez *et al*., 2019). Linear algorithms, such as Linear Regression assume a straightforward relationship between input features and output variables (variable to be predicted). They are simple, computationally efficient and easy to interpret, but often fail to capture complex patterns in data. Conversely, non-linear algorithms can model complex patterns and interactions in data more effectively (López & Arboleya, 2022). Both linear and non-linear algorithms have been extensively analysed for crop yield prediction in recent years. For instance, a study (Ekanayake *et al*., 2021) used Linear Regression and many other machine learning models to predict paddy yield. Similarly, (Han *et al*., 2023) applied Linear Regression to understand meteorological impacts on apple yield and prediction. From the perspective of agricultural economic management, Yang *et al*., (2024) predicted corn yield using Random Forest that is a non-linear tree-based algorithm. Similarly, Extra Trees regression, another tree-based method, was reported as the best-performing model for cotton yield prediction using weather data (Sudhamathi & Perumal, 2024).

Despite the demonstrated capabilities of linear and non-linear machine learning models, their potential for solving the oil palm yield prediction problem remains underexplored. The oil palm industry faces challenges such as fluctuating weather conditions and the need for sustainable practices to mitigate environmental impact. Accurate yield prediction is required to optimise production, improve field management, ensure proper crop handling, facilitate trade agreements, support economic planning and enhance food security (An-Vo *et al*., 2021). Advanced machine learning techniques can address these challenges by providing reliable forecasts of fruit production (Khan *et al*., 2021; Parveez *et al*., 2020). Recent advancements in machine learning have improved crop yield prediction, particularly in oil palm plantations. Prior studies have extensively explored linear models such as Linear Regression due to their interpretability (Ang *et al*., 2022; Oettli *et al*., 2018; Watson-Hernández *et al*., 2022), while non-linear ensemble methods like Extra Trees have shown better predictive accuracy but are often criticised for their limited interpretability (Jamshidi *et al*., 2024; Khan, 2023; Mohd Nain *et al*., 2022). To date, limited research has directly compared these two modelling approaches for oil palm yield forecasting. This study addresses this gap by systematically comparing the performance of a linear model (Linear Regression) with non-linear tree-based models such as Extra Trees, Gradient Boosting, and Random Forest for oil palm yield prediction. The comparison is based on historical fresh fruit bunch (FFB) yields and weather data from oil palm plantations in Pahang, Malaysia, covering the years 1986-2020. The study emphasises the importance of data preprocessing, feature selection, and model evaluation in developing reliable predictive systems. The findings aim to assist stakeholders in selecting optimal modelling approaches for sustainable farming practices and informed decision-making within the oil palm industry. Insights into the models and oil palm responses to environmental variations further highlight the impacts of irregular weather patterns on crop yield. The article is structured as follows: The Materials and Methods section describes the study site, data collection, and preprocessing. The Results and Discussion section presents the performance evaluation and comparison of models, while the Conclusion summarises the work and provides the key findings with their implications.

MATERIALS AND METHODS

This study aimed to use data-driven supervised machine learning for predictive modelling of oil palm yield. The investigation began with the identification of a comprehensive dataset, the geographical study location, and the temporal scope of the data. Linear Regression was selected as the baseline model due to its simplicity and interpretability. In contrast, among the various non-linear algorithms, three tree-based models were selected due to their suitability to the size and complexity of the dataset. Details regarding the study area, data acquisition, preprocessing steps, and model selection criteria are provided in the following subsections.

Data and Location

Pahang, situated on the east coast of Peninsular Malaysia at 4°11′10″N and 104°03′45″E, covers an area of 35,965 km^2. It has a hot, humid tropical climate that can be described as a rainforest tropical climate. The oil palm sector in Pahang is at risk from climate change-induced crop stagnation (Rhebergen *et al.*, 2016). This is why Pahang is chosen as the study area. *Figure 1* illustrates the geographic location of Pahang in relation to Malaysia.

Predicting oil palm yields with precision requires a robust and technically sound approach that integrates multiple factors, including crop growth patterns, soil conditions and weather dynamics. To achieve this, multisource time-series data from remote sensing platforms, soil sensors, and meteorological stations are collected and combined into a unified framework. The data from different sources is synchronised by aligning it within a consistent temporal and spatial resolution, ensuring compatibility for accurate yield predictions.

For this study, a multivariate dataset is obtained from Malaysian Palm Oil Board (MPOB), Jabatan Meteorologi Malaysia (MET), NASA Data Access Viewer and Soil Grids. The data included 420 monthly average values of 17 variables: 'Oil palm yield', 'temperature range', 'surface pressure', 'earth skin temperature', 'specific humidity', 'precipitation', 'cloud amount', 'rain days', 'wind speed', 'rainfall', 'solar irradiance', 'relative humidity', 'profile soil moisture', 'surface soil wetness', 'root zero soil wetness', 'minimum temperature' and 'maximum temperature'. It should be noted that each data variable within an identical timeframe of monthly values (throughout January 1986-December 2020) was integrated into a single data frame. Moreover, raw data needed to be prepared before training machine learning models. Data pre-reprocessing contributed to improving data quality by treating inconsistencies in the data. The step-by-step preprocessing procedure is explained in the following section.

Data Preprocessing

Raw data is processed before building and training a machine learning model. The primary objective of data preprocessing is to enhance the overall performance and reliability of the machine learning algorithm by transforming raw data into meaningful information. For this study, outliers (inconsistent values in data) are statistically treated using the flooring and capping method, an established technique to deal with outliers in real-world small data. Then, data is aggregated using a rolling mean to capture annual trends from each feature.

The process of applying a rolling window on features is given in *Figure 2*, where X refers to the original features, Y is the grouping of each feature value in a predefined window, *d* is the window size and D is the transformed data obtained after aggregation.

The features in multivariate data are measured on a variety of scales that may drastically affect the performance of machine learning algorithms.

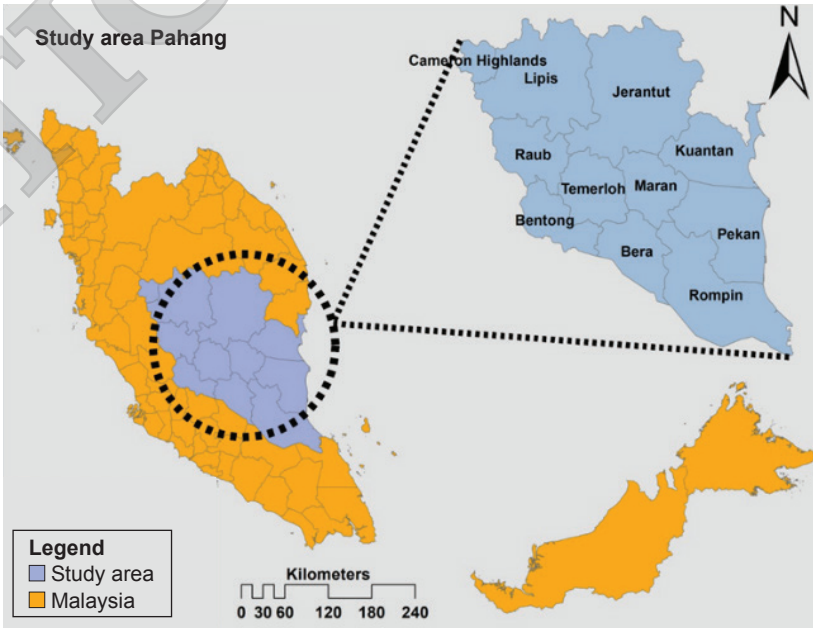

Figure 1. Study area: Pahang state, Malaysia.

Without feature scaling, outcomes are affected by generating biased results due to high magnitudes in certain features. This issue is addressed statistically to fix the range of independent features. Numerical stability is created in all features by setting similar magnitudes. For feature scaling, the minimum and maximum scaling is performed using the following Equation (1).

$$X\_scaled = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

Finally, prepared data is divided into training, validation, and testing sets. First, 10% of the clean data is kept unseen for testing purposes. The remaining 90% of the data is scaled and divided into 80:20 ratios for model training and validation using a 10-fold cross-validation method. A detailed description of the data splitting procedure is given in *Figure 3*.

Training, Evaluation, and Comparison of Machine Learning Models

From the prepared data set, which consisted of 80% training set and 20% validation set, only the training set was used to train machine learning models. Linear Regression and the tree-based models represent a broad range of linear and non-linear models. These models were selected based on the characteristics, size and dimensions of the dataset. Linear Regression was chosen for its simplicity and interpretability, while tree-based models effectively capture non-linear relationships in moderately sized datasets. Support vector machine (SVM), though effective for predictions, requires extensive preprocessing for high-dimensional data (Benhar *et al.*, 2020). Artificial neural networks (ANNs) were excluded due to their higher data requirements. The selected models for training and comparison are detailed below. The schematic diagrams of the selected models are also given in *Figure 4*.

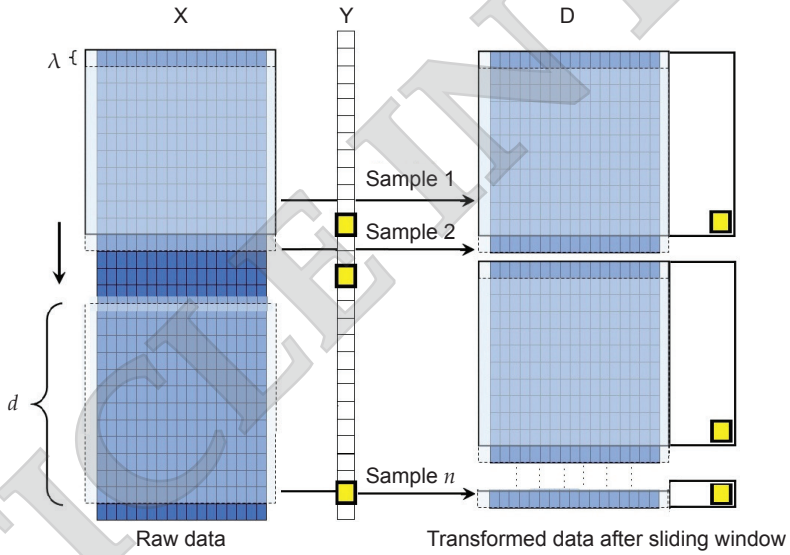

Figure 2. Data aggregation using the rolling window technique.

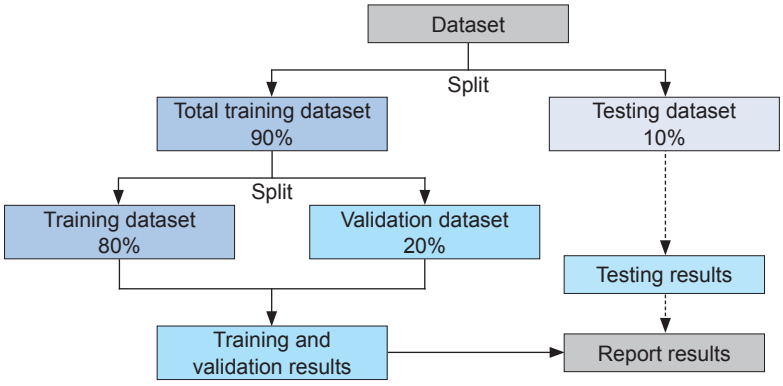

Figure 3. Data splitting mechanism.

Linear Regression. Linear regression is a popular supervised machine learning method for crop yield prediction. Linear regression models the linear relationship between one or more independent variables (influencing factors) and a dependent variable (crop yield). It can be used to predict crop yield based on weather data, soil characteristics and other factors.

Random Forest Regression. Random Forest is an ensemble method that builds multiple decision trees using random subsets of data and features. It combines their outputs by majority voting or averaging to improve accuracy and reduce overfitting. The key difference between Random Forest and Extra Trees is that Random Forest selects the best split based on a random subset of features, while Extra Trees chooses split points randomly (Kwak *et al.*, 2022).

Gradient Boosting Regression. Gradient Boosting is a decision tree algorithm that builds models sequentially. In each iteration, a new tree is trained to predict the residual errors of the combined predictions from previous trees. The output of the new tree is added to improve the overall prediction. This process continues for a set number of iterations or until a stopping criterion is met (Iniyan & Jebakumar, 2022).

Extra Trees Regression. Extra Trees regression, or Extremely Randomised Trees, is an ensemble learning algorithm designed for high-performance regression tasks. It extends the traditional decision tree approach by introducing greater randomisation during the tree construction process, which helps reduce overfitting and enhances model generalization for complex agricultural data (Jamshidi *et al.*, 2024). Unlike Random Forest, which relies on bootstrapped samples, Extra Trees builds each tree using the entire dataset but introduces randomness by selecting split points within features at random (Salman *et al.*, 2024). With this additional layer of randomness, Extra Trees achieves better computational efficiency for datasets with non-linear and complex relationships that often exist in real-world time series data. The algorithm is particularly effective in scenarios with moderate to large feature sets and is suitable for environmental modelling and yield prediction tasks. The hyperparameters for Extra Trees were optimised using a Grid Search approach to select the best-performing configuration.

The prepared dataset was systematically employed to train selected linear and non-linear machine learning regression models in an identical manner. Through this comprehensive trialling process, the outcomes in the form of predictions were obtained. The performance of the models was

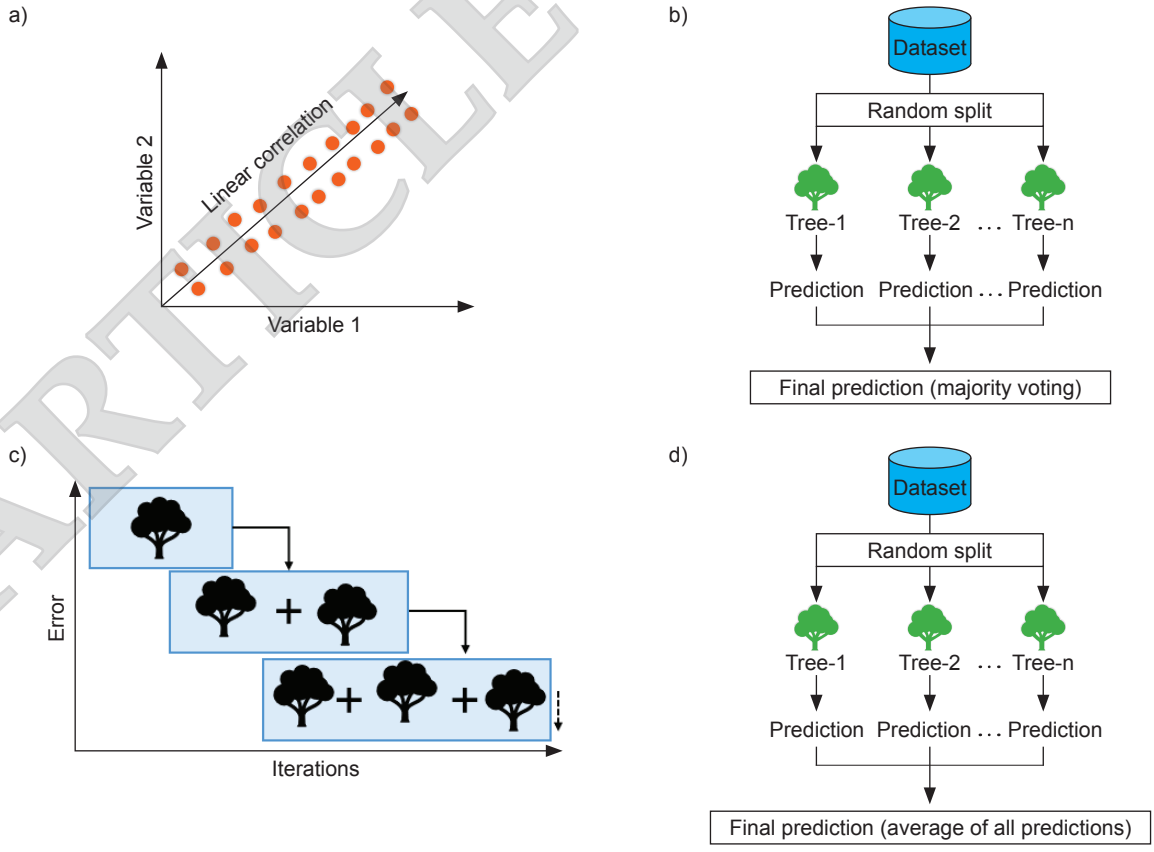

Figure 4. Schematic diagram of (a) Linear regression, (b)Random Forest, (c) Gradient Boosting and (d) Extra Trees.

evaluated using mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2) as key metrics.

RESULTS AND DISCUSSION

This section includes the yield trends throughout the study period (1986-2020) and details the correlation among data variables to understand the impact of different environmental factors on oil palm yield. It also includes prediction outcomes obtained after training machine learning models, Linear regression, Random Forest, Gradient Boosting and Extra Trees regression. Furthermore, this section details an assessment of the impact of different parameters on predictions of the best-performing model.

Yield Trends, Data Variables and Correlations Between Features

The annual trends of oil palm yield in Pahang recorded on a monthly basis from 1986-2020 are presented in *Figure 5*. Throughout the study period, no upward trends exist in yield records. Instead, the lowest yield was observed in 2017. Surprisingly, the yield could not increase despite the technological advancements in terms of quality seeds, modern fertilisers and developed crop protection through weed detection and disease control (Tian *et al.*, 2020). Oil palm yield is influenced by several biotic (pests, insects, disease, weeds, *etc.*) and abiotic (weather, soil, fertilisers) factors. The reasons behind fluctuating yield can be assessed through yield gap analysis, as discussed in (Khan *et al.*, 2022a). Oil palm yield trends contradict the research (Van Ittersum *et al.*, 2013) suggesting crop yields need to increase to meet rising global food demand, oil palm production is not increasing.

Environmental factors can significantly influence crop yields, either positively or negatively, depending on the crop type and its environmental requirements. Thorough examination of these impacts is important for selecting suitable sites and mitigating negative environmental effects on oil palm. A correlation analysis of all data variables was conducted to investigate these relationships, as presented in *Figure 6*. This analysis highlights the detrimental effects of high wind speeds and temperature fluctuations represented by the temperature range. Additionally, rainfall and solar irradiance exhibit low positive correlation values because of their conditional correlations with oil palm yield. Conditional correlation means their effects on oil palm productivity can be positive or negative, depending on their levels. The correlation analysis aligns with findings from previous studies, confirming the significant influence of key environmental and agronomic factors on oil palm yield prediction (Khan *et al.*, 2022a; Monzon *et al.*, 2023). Correlations between each variable and yield, if taken as environmental impact on yield variations, are beneficial for understanding the reasons behind yield reductions through yield gap analysis. On the other hand, these correlations from the machine learning perspective reflect the importance of features for models describing how informative a feature can be during the training process. Machine learning models capture the correlations and mathematically map historical values of input variables with corresponding yield values. Afterward, future outcomes are predicted based on dependencies among features learned from previous trends.

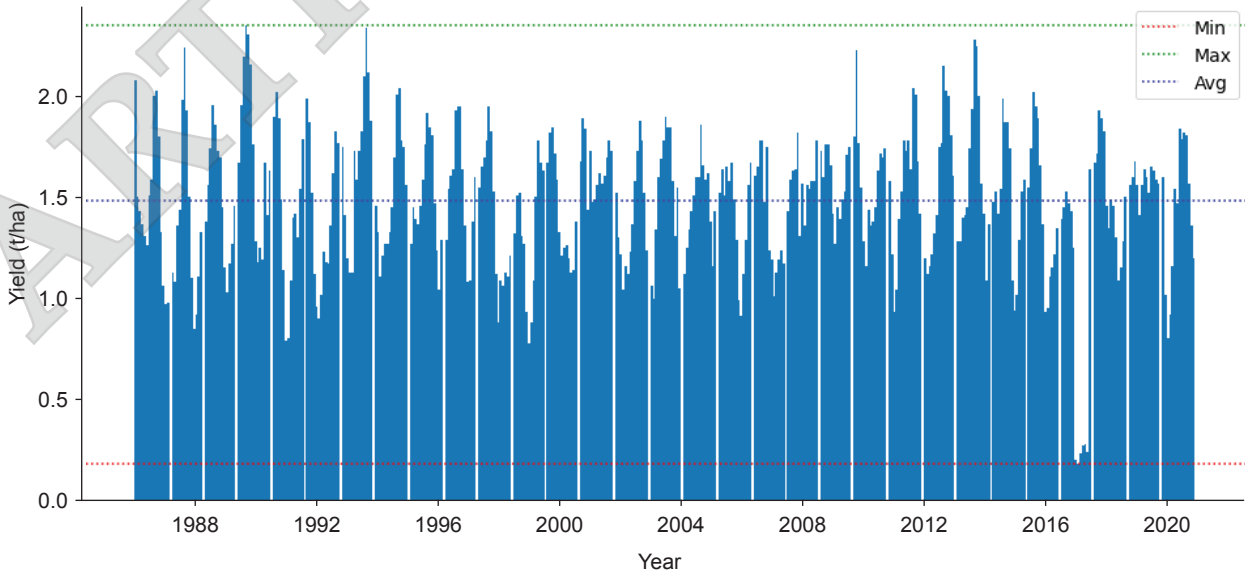

Figure 5. Oil palm yield trends from 1986-2020 in Pahang.

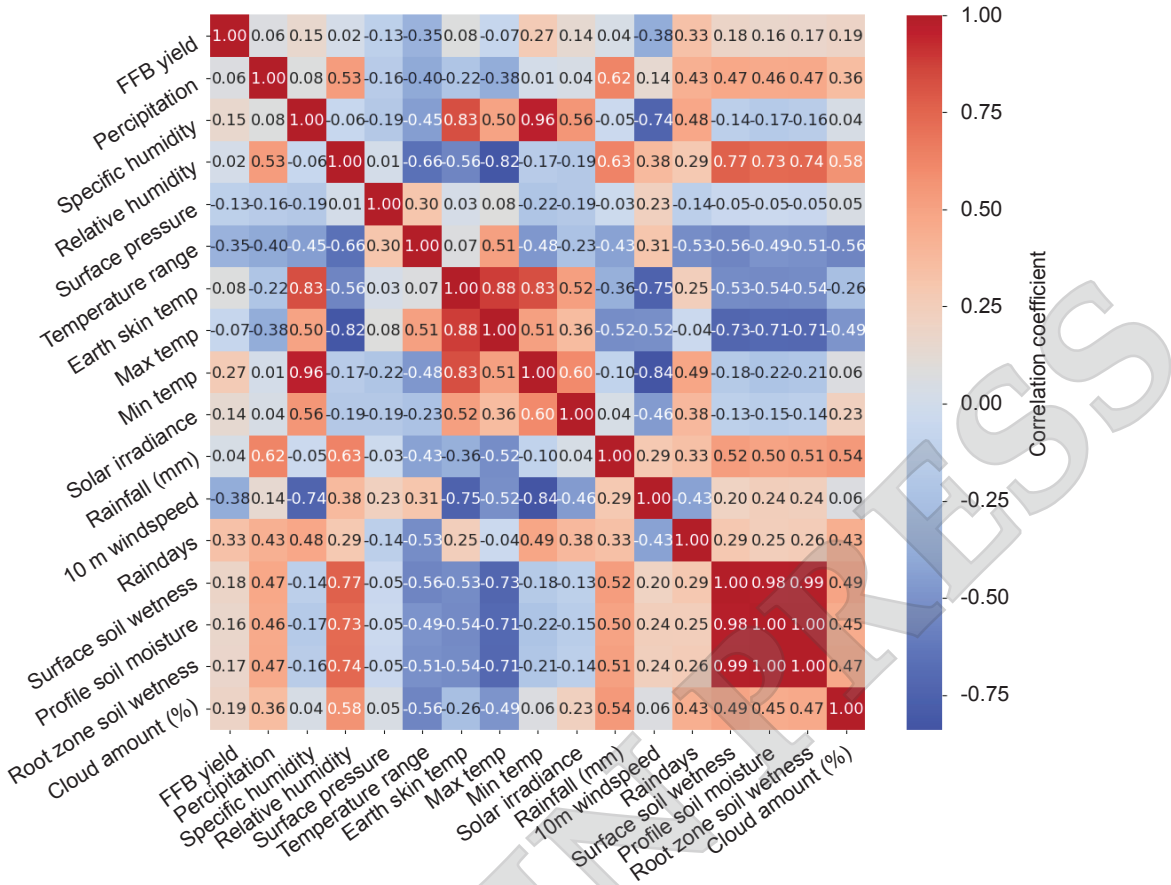

Figure 6. Features and correlations in data.

Prediction of Oil Palm Yield by Machine Learning Regression

Several standard statistical evaluation metrics were used to assess the prediction accuracy of the machine learning models, including MSE, MAE, MAPE, Root Mean Squared Error (RMSE) and R^2. The performance results of Linear Regression, Random Forest, Gradient Boosting and Extra Trees are summarised in *Table 1*. Among these metrics, MSE, MAE, MAPE and RMSE quantify the prediction error by measuring the difference between actual and predicted values. Lower values of these error metrics indicate better model performance, signifying reduced prediction errors. Conversely, a higher validation and testing R^2 value reflects better model accuracy. It represents the proportion of variance in the target variable based on input variables explained by the model.

The results presented in *Table 1* clearly demonstrate that the Extra Trees model achieved the best performance among all evaluated models, with an R^2 value of 88% on the validation dataset and the lowest error metrics. It was followed by Random Forest (R^2 = 79%) and Gradient Boosting (R^2 = 78%), both showing strong predictive capabilities. In contrast, Linear Regression achieved a significantly lower R^2 of 41%, indicating limited suitability for this prediction task. The high R^2 of the Extra Trees model highlights its effectiveness in capturing yield variability, making it a valuable decision-support tool for agricultural planners in Pahang state. These predictions can aid in capacity planning and proactive environmental management, helping to mitigate negative impacts on future oil palm yield through timely interventions. Although the model explains 88% of the variance, the remaining 12% unexplained variability may be attributed to missing factors particularly biotic influences that were not included in the dataset. Incorporating such factors could further enhance predictive accuracy. These outcomes are consistent with a prior study (Jamshidi *et al.*, 2024) that reported superior performance of Extra Trees on oil palm plantation data. However, the Extra Trees model has some limitations, such as computational complexity, which grows with the size of the dataset and depth of trees. Additionally, while non-linear models excel in capturing complex relationships in data, they risk overfitting in the presence of limited or noisy data (Sudhamathi & Perumal, 2024). These trade-offs must be considered when selecting large-scale, real-world agricultural application models.

TABLE 1. STATISTICAL EVALUATION METRICS OF SELECTED MODELS

Model	MSE	MAE	MAPE	RMSE	Validation R^2
Linear Regression	0.0050	0.0583	0.0394	0.0710	41%
Gradient Boosting	0.0399	0.3210	0.0216	0.0399	78%
Random Forest	0.0015	0.0310	0.0208	0.0389	79%
Extra Trees	0.0014	0.0246	0.0165	0.0350	88%

On the other hand, the lower R^2 value of 41% for the Linear Regression model highlights its inherent limitations. Linear models are generally suited to datasets where relationships between variables are predominantly linear. However, in this case, many of the influencing factors particularly weather variables exhibit non-linear patterns. As a result, Linear Regression fails to adequately capture the complexity of these relationships, leading to underestimation of key non-linear effects on yield variability. This structural mismatch contributes to the model's higher prediction errors and reduced effectiveness compared to more flexible, non-linear approaches. In the literature, several statistical techniques, such as logarithmic, polynomial, or Box-Cox transformations (Demir & Sahin, 2024), have been applied to convert non-linear data into a linear form to make it compatible with linear models. While these transformations can simplify modelling, they may also distort the original data patterns and result in the loss of critical information. Such approaches may be acceptable in scenarios where the simplicity of the model is prioritised over data integrity. However, for complex applications like crop yield prediction, where environmental variables exhibit inherent non-linearity, linear methods are often inadequate. Accurate modelling in agricultural contexts requires alignment between model structure and data complexity to ensure meaningful insights and effective decision-making. This is particularly important when predictions inform large-scale resource allocation, where small errors may lead to significant inefficiencies. Therefore, in this study, we adopted non-linear machine learning models capable of capturing complex relationships within the data, thus ensuring better predictive performance and more reliable decision-making in yield forecasting. Results showed that non-linear models demonstrated superior predictive performance as compared to Linear Regression. Despite being trained on the same dataset, Linear Regression resulted in significantly higher prediction errors. Among tree-based models, the Extra Trees model outperformed others by providing more accurate and reliable oil palm yield predictions. The experiment indicates the importance of selecting machine learning models that are well-matched to the underlying characteristics of the dataset. This observation can be generalised to choose suitable machine learning models based on data characteristics.

Performance Comparison of Models on Validation Data

Besides the performance metrics, a two-fold in-depth analysis of the fitting process of each model is observed to examine the best-fit line of each model on 1) validation data and 2) unseen testing data. *Figure 7* highlights the trend between each predicted value and actual value for Extra Trees and Linear Regression, respectively. The actual values are on the X-axis of each figure, and predicted values are plotted on the Y-axis. Actual values and predicted values are denoted by y and $\hat{y}$, respectively. To evaluate how well the models follow data trends, we plotted an ideal identity dashed line in grey and the best fit dashed line of the models in black. The closer the best-fit line of a model is to the identity line, the more accurate its predictions are. Deviations from this line indicate prediction errors. Detailed performance visualisations are shown in *Figure 7a* and *7b* for the best-performing model (Extra Trees) and the least accurate model (Linear Regression), respectively. These were selected to illustrate the full performance spectrum and provide a clear contrast in model capabilities.

Even though the evaluation metrics and error plots are a reliable way to compare the prediction accuracy and generalisation power of machine learning models, there are situations when a model that performs well on validation data suffers from overfitting. It usually happens when validation data is leaked, and models accidentally get exposed to validation data during data shuffling. In that situation, a model may exhibit good accuracy on validation data in terms of evaluation metrics and error plots, but it performs poorly on new, unseen data. This overfitting is an undesirable outcome, particularly when prediction models are trained for industrial and economic applications, as incorrect predictions may lead to poor decisions and financial losses.

Performance Comparison of Machine Learning Models on Unseen Data

A two-step validation process is employed to ensure the reliability of machine learning models. Once the models were validated on the validation dataset, predictions were made on an additional 10% of testing data, which had never been used

during the training or initial validation phases. This unseen testing data is used further to examine the accuracy of models on entirely new data. The proposed multilayer performance evaluation approach enhanced the confidence in the predictive capabilities of the non-linear models and their applicability to real-world scenarios. *Figure 8* and *9* show the prediction of oil palm yield by Extra Trees and Linear Regression models, respectively. The accuracy of the models is evaluated by plotting actual values on the X-axis and predicted values on the Y-axis. A red trend line represents the ideal perfect prediction, while blue dots illustrate the individual differences between actual and predicted values. The clustering of blue dots around the red line demonstrates the predictive performance of models on new unseen data. A high density of dots near the red fitting line suggests a small magnitude of errors that indicate a high level of accuracy. A wider spread indicates larger prediction errors, as seen in the Linear Regression case. One random sample of data points is highlighted on the figures with red circles and actual and predicted values for a specific instance. These figures show that the non-linear models exhibit higher accuracy, displaying a tighter clustering of dots around the red trend line than the Linear Regression model. The consistency of outcomes between the test data and the validation data demonstrates the robustness of Extra Trees in the training process and the reliability of predictions on test data. By maintaining similar performance levels on the unseen validation data, it is ensured that the models are not overfitting and can reliably generalise to new, unobserved data. The reason for the poor performance of Linear Regression on the given data is that it could not capture the inherent non-linearity in agricultural data. The superior performance of Extra Trees over Linear Regression is consistent with previous research, where tree-based ensemble methods achieved higher accuracy in agricultural yield prediction (Chen *et al*., 2022; Sudhamathi & Perumal, 2024).

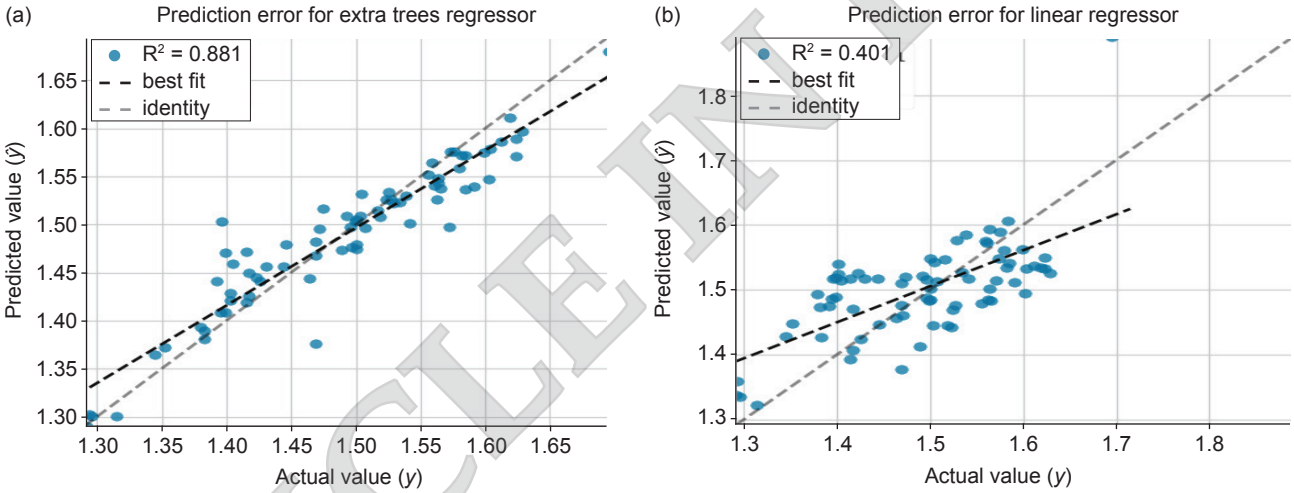

Figure 7. Prediction error of (a) Extra Trees and (b) Linear Regression on validation data.

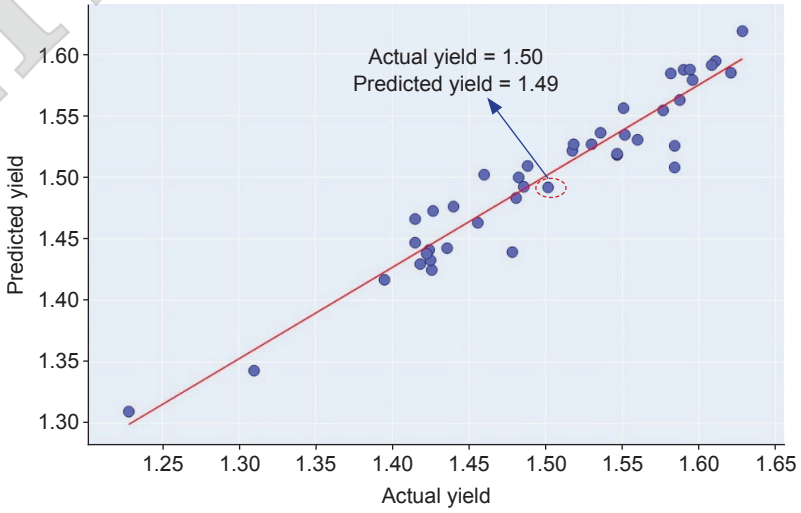

Figure 8. Performance of Extra Trees on unseen data.

Machine learning models learn from data containing various features. It is important to understand the contribution of each feature to the predicted oil palm yield. Since the Extra Trees model demonstrated higher accuracy, examining its learning process can explain the significance of various features in making accurate predictions. An additive explanation plot for the Extra Trees model is provided in *Figure 10* to understand the contributions of individual features to the oil palm yield predictions. The weights given to each feature value by the model are provided on the X-axis, while features on the Y-axis are ranked from high to low based on their impact on oil palm yield predictions. The plot shows the positive and negative values plotted in blue and red colours, respectively, to visualise the SHAP value of every data point.

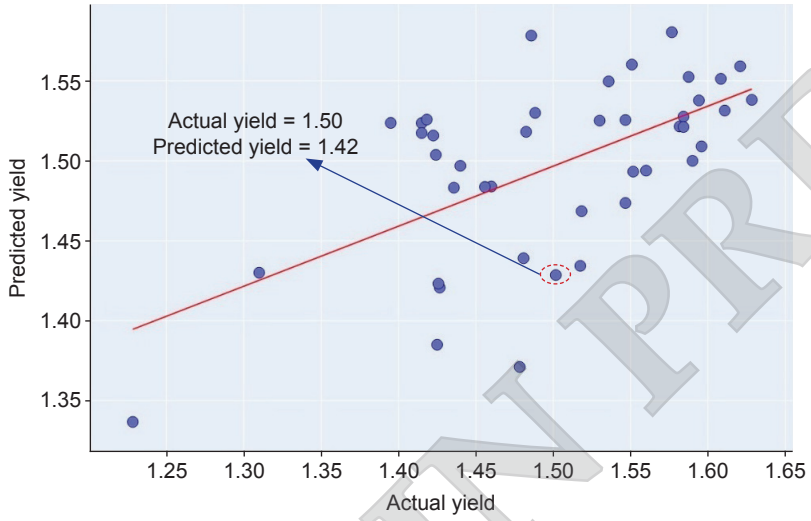

Figure 9. Performance of Linear Regression on unseen data.

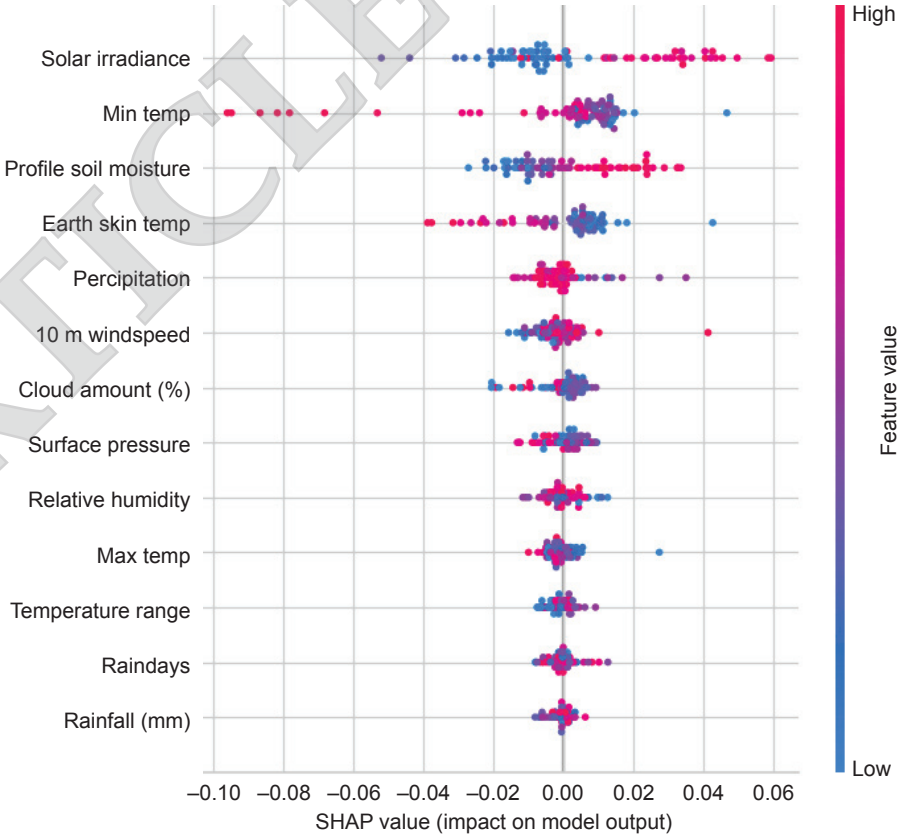

Figure 10. Impact of features on predictions of Extra Trees.

The feature importance plot provided a detailed view of how each feature contributed to the predictions of Extra Trees when used as an input feature. Moreover, the role of environmental factors on oil palm yield is analysed. As observed, solar irradiance is the most significant predictor of oil palm yield. The model shows higher accuracy when solar irradiance values are high. It reflects the positive contribution of solar irradiance to the model in explaining yield variability. Similarly, in real-world contexts, increased sunlight availability enhances yield, while lower irradiance adversely impacts both predictions and actual oil palm production. Conversely, minimum temperature demonstrated that lower values positively influenced and proved more informative than higher values. Extra Trees captured more trends from low values of minimum temperature, perhaps due to strong responses of oil palm crops to cold weather (Jamshidi *et al.*, 2024). This is because the oil palm crop is sensitive to cold and shows significant yield variations in low temperatures (Li *et al.*, 2019). Similarly, profile soil moisture, earth skin temperatures and precipitation are among the top five useful features for the model. Through a deeper understanding of feature importance, this analysis can improve crop management (Monzon *et al.*, 2023) and yield prediction strategies. Furthermore, it guides future studies in optimising data selection for efficient predictive modelling and other agricultural practices.

The study are based on real-world data, incorporating various scenarios, including drought, heavy rainfall, temperature variations, sunlight, and soil moisture conditions. Instead of analysing each scenario separately, the model captures these effects jointly, reflecting real-world conditions. This approach aligns with the study's objective of comparing model performance in a practical setting rather than a controlled environment. This study will be expanded to explore different scenarios, including weather variations, regional productivity differences and yield distributions.

CONCLUSION

The oil palm agriculture sector stands to benefit significantly from the adoption of advanced machine learning techniques to tackle challenges posed by climate change, enhance productivity, optimise resource utilisation, and promote sustainable practices. At the core of this transformation lies the need for reliable yield prediction, which remains a complex task due to the integration and effective utilisation of diverse agro-environmental data. This study conducted a comparative evaluation of linear and non-linear machine learning models for oil palm yield prediction using historical weather and soil moisture data from Pahang, Malaysia. Linear Regression was used as a baseline, while three tree-based non-linear models including Extra Trees, Random Forest and Gradient Boosting were assessed for their ability to capture complex environmental interactions. The findings clearly show that non-linear models significantly outperformed the linear approach. The Extra Trees model achieved the highest performance, with an R^2 of 88% on validation data, closely followed by Random Forest (79%) and Gradient Boosting (78%). These models also exhibited lower prediction errors, further demonstrating their effectiveness in modeling complex, non-linear yield influencing patterns. In contrast, Linear Regression, with an R^2 of only 41%, failed to adequately model the variability in yield due to its inability to capture non-linear relationships inherent in environmental data. The close alignment between actual and predicted yield using the Extra Trees model highlights its reliability. Minor deviations can likely be attributed to unaccounted abiotic factors (*e.g.,* soil fertility, management practices) and biotic influences (*e.g.,* pests, weeds, disease outbreaks) not included in the dataset. Nonetheless, the study confirms the suitability of weather and soil moisture variables for training machine learning models in yield forecasting and reinforces the advantage of non-linear approaches in agricultural prediction tasks. By integrating long-term historical yield records with diverse environmental parameters, this study contributes toward addressing the challenges of data scarcity and multisource data incompatibility. Following extensive data preprocessing, the Extra Trees model proved capable of delivering reliable yield forecasts, supporting resource allocation, capacity planning, sustainable intensification and strategic decision-making in oil palm agriculture.

The study also reveals that optimal sunlight positively impacts oil palm yield in Pahang, while high winds, cold temperatures, and water limitations are detrimental. These insights are vital for improving crop management practices. The outcomes can be informative for future research in data, features and model selection for predictive modelling of different crops. The proposed forecasting procedure is generic and can be applied to other datasets with different variables or to crops other than oil palm. This study also benefits researchers and developers in designing smart agricultural tools.

These findings highlight the potential of ensemble methods for improving agricultural forecasting. The non-linear complex models can enhance farm productivity, yield monitoring

and decision-making processes. The data-driven approach can lead to more innovations in farm business models to improve the environmental and economic performance of the agricultural sector. Future research will focus on refining the models with broader datasets containing quantified biotic and abiotic factors. Moreover, there is strong potential to improve prediction accuracy by exploring advanced machine learning and deep learning techniques, such as powerful neural networks. This study was carried out at the state level, covering large-scale oil palm plantations, and the proposed methods are designed to be spatially adaptable and transferable. They can be applied across different scales, *i.e.*, from entire plantations down to individual plots for more targeted insights. With the integration of fine-grained data collected at the individual tree level, these methods can be further refined to support high-precision yield estimation and localised decision-making. Future research can focus on developing hybrid models, incorporating deep learning approaches, and integrating remote sensing data to improve both the accuracy and practical relevance of yield predictions. In the long run, adopting machine learning in oil palm cultivation has the potential to empower farmers and guide policymakers toward more sustainable and data-driven agricultural practices.

ACKNOWLEDGEMENT

This work was funded by the School of Industrial Technology, Universiti Sains Malaysia, under Grant 203.PTEKIND.6777007. The authors would like to thank the Ministry of Higher Education, Malaysia, for their support through the Long Term Research Grant Scheme. We would also like to thank the MPOB and the MET for providing the essential data for this research.

REFERENCES

An-Vo, D. A., Radanielson, A. M., Mushtaq, S., Reardon-Smith, K., & Hewitt, C. (2021). A framework for assessing the value of seasonal climate forecasting in key agricultural decisions. *Climate Services*, *22*, 100234. https://doi.org/10.1016/j.cliser.2021.100234

Ang, Y., Shafri, H. Z. M., Lee, Y. P., Bakar, S. A., Abidin, H., Junaidi, M. U. U. M., Hashim, S. J., Che'Ya, N. N., Hassan, M. R., Lim, H. S., Abdullah, R., Yusup, Y., Muhammad, S. A., Teh, S. Y., & Samad, M. N. (2022). Oil palm yield prediction across blocks from multi-source data using machine learning and deep learning. *Earth Science Informatics*, *15*(4), 2349–2367. https://doi.org/10.1007/s12145-022-00882-9

Benhar, H., Idri, A., & Fernández-Alemán, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, *195*, 105635. https://doi.org/10.1016/j.cmpb.2020.105635

Bharadiya, J. P., Tzenios, N. T., & Reddy, M. (2023). Predicting crop yield using deep learning and remote sensing. *Journal of Engineering Research and Reports*, *24*(12), 29–44. https://doi.org/10.9734/jerr/2023/v24i12858

Chen, R., Zhang, C., Xu, B., Zhu, Y., Zhao, F., Han, S., Yang, G., & Yang, H. (2022). Predicting individual apple tree yield using UAV multi-source remote sensing data and ensemble learning. *Computers and Electronics in Agriculture*, *201*, 107275. https://doi.org/10.1016/j.compag.2022.107275

Demir, S., & Sahin, E. K. (2024). The effectiveness of data pre-processing methods on the performance of machine learning techniques using RF, SVR, Cubist and SGB: A study on undrained shear strength prediction. *Stochastic Environmental Research and Risk Assessment*, *38*(8), 3273–3290. https://doi.org/10.1007/s00477-024-02745-9

Ekanayake, P., Rankothge, W., Weliwatta, R., & Jayasinghe, J. W. (2021). Machine learning modelling of the relationship between weather and paddy yield in Sri Lanka. *Journal of Mathematics*, *2021*, 1–14. https://doi.org/10.1155/2021/9941899

Gebresenbet, G., Bosona, T., Patterson, D., Persson, H., Fischer, B., Mandaluniz, N., Chirici, G., Zacepins, A., Komasilovs, V., Pitulac, T., & Nasirahmadi, A. (2023). A concept for application of integrated digital technologies to enhance future smart agricultural systems. *Smart Agricultural Technology*, *5*, 100255. https://doi.org/10.1016/j.atech.2023.100255

Han, X., Chang, L., Wang, N., Kong, W., & Wang, C. (2023). Effects of meteorological factors on apple yield based on multilinear regression analysis: A case study of Yantai Area, China. *Atmosphere*, *14*(1), 183. https://doi.org/10.3390/atmos14010183

Iniyan, S., & Jebakumar, R. (2022). Mutual information feature selection (MIFS) based crop yield prediction on corn and soybean crops using multilayer stacked ensemble regression

(MSER). *Wireless Personal Communications*, *126*(3), 1935–1964. https://doi.org/10.1007/s11277-021-08712-9

Jamshidi, E. J., Yusup, Y., Hooy, C. W., Kamaruddin, M. A., Hassan, H. M., Muhammad, S. A., Shafri, H. Z. M., Then, K. H., Norizan, M. S., & Tan, C. C. (2024). Predicting oil palm yield using a comprehensive agronomy dataset and 17 machine learning and deep learning models. *Ecological Informatics*, *81*, 102595. https://doi.org/10.1016/j.ecoinf.2024.102595

Javaid, M., Haleem, A., Khan, I. H., & Suman, R. (2023). Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, *2*(1), 15–30. https://doi.org/10.1016/j.aac.2022.10.001

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, *10*, 01750.https://doi.org/10.3389/fpls.2019.01750

Khan, N. (2023). *Prediction of oil palm yield for smallholders estates in tropical region using extra trees method* [Doctoral dissertation]. Universiti Sains Malaysia.

Khan, N., Kamaruddin, M. A., Sheikh, U. U., Yusup, Y., & Bakht, M. P. (2021). Oil palm and machine learning: Reviewing one decade of ideas, innovations, applications, and gaps. *Agriculture*, *11*(9), 832. https://doi.org/10.3390/agriculture11090832

Khan, N., Kamaruddin, M. A., Sheikh, U. U., Yusup, Y., & Bakht, M. P. (2022a). Oil palm yield gap prediction using machine learning: A proof of concepts. *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*. https://doi.org/10.1109/macs56771.2022.10022394

Khan, N., Kamaruddin, M. A., Sheikh, U. U., Zawawi, M. H., Yusup, Y., Bakht, M. P., & Noor, N. M. (2022b). Prediction of oil palm yield using machine learning in the perspective of fluctuating weather and soil moisture conditions: Evaluation of a generic workflow. *Plants*, *11*(13), 1697. https://doi.org/10.3390/plants11131697

Kwak, S., Kim, J., Ding, H., Xu, X., Chen, R., Guo, J., & Fu, H. (2022). Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model. *Journal of Materials Research and Technology*, *18*, 520–530. https://doi.org/10.1016/j.jmrt.2022.02.108

Li, J., Yang, Y., Iqbal, A., Qadri, R., Shi, P., Wang, Y., Wu, Y., Fan, H., & Wu, G. (2019). Correlation analysis of cold-related gene expression with physiological and biochemical indicators under cold stress in oil palm. *PLoS ONE*, *14*(11), e0225768. https://doi.org/10.1371/journal.pone.0225768

López, G., & Arboleya, P. (2022). Short-term wind speed forecasting over complex terrain using linear regression models and multivariable LSTM and NARX networks in the Andes Mountains, Ecuador. *Renewable Energy*, *183*, 351–368. https://doi.org/10.1016/j.renene.2021.10.070

Mendez, K. M., Reinke, S. N., & Broadhurst, D. I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, *15*(12). https://doi.org/10.1007/s11306-019-1612-4

Mohd Nain, F. N., Malim, N. H. A. H., Abdullah, R., Rahim, M. F. A., Mokhtar, M. A. A., & Fauzi, N. S. M. (2022). A review of an artificial intelligence framework for identifying the most effective palm oil prediction. *Algorithms*, *15*(6), 218. https://doi.org/10.3390/a15060218

Monzon, J. P., Lim, Y. L., Tenorio, F. A., Farrasati, R., Pradiko, I., Sugianto, H., Donough, C. R., Edreira, J. I. R., Rahutomo, S., Agus, F., Slingerland, M. A., Zijlstra, M., Saleh, S., Nashr, F., Nurdwiansyah, D., Ulfaria, N., Winarni, N. L., Zulhakim, N., & Grassini, P. (2023). Agronomy explains large yield gaps in smallholder oil palm fields. *Agricultural Systems*, *210*, 103689. https://doi.org/10.1016/j.agsy.2023.103689

Oettli, P., Behera, S. K., & Yamagata, T. (2018). Climate based predictability of oil palm tree yield in Malaysia. *Scientific Reports*, *8*(1), 2271. https://doi.org/10.1038/s41598-018-20298-0

Parveez, G. K. A., Hishamuddin, E., Loh, S. K., Ong-Abdullah, M., Salleh, K. M., Bidin, M., Sundram, S., Hasan, Z. A. A. & Idris, Z. (2020). Oil palm economic performance in Malaysia and R&D progress in 2019. *Journal of Oil Palm Research, 32*(2), 159–190. https://doi.org/10.21894/jopr.2020.0032

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE Access*, *9*, 63406–63439. https://doi.org/10.1109/access.2021.3075159

Rhebergen, T., Fairhurst, T., Zingore, S., Fisher, M., Oberthür, T., & Whitbread, A. (2016). Climate, soil and land-use based land suitability evaluation for oil palm production in Ghana. *European Journal of Agronomy*, *81*, 1–14. https://doi.org/10.1016/j.eja.2016.08.004

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. https://doi.org/10.58496/bjml/2024/007

Sudhamathi, T., & Perumal, K. (2024). Ensemble regression based extra tree regressor for hybrid crop yield prediction system. *Measurement Sensors*, *35*, 101277. https://doi.org/10.1016/j.measen.2024.101277

Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation – A review. *Information Processing in Agriculture*, *7*(1), 1–19. https://doi.org/10.1016/j.inpa.2019.09.006

Van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., & Hochman, Z. (2013). Yield gap analysis with local to global relevance – A review. *Field Crops Research*, *143*, 4–17. https://doi.org/10.1016/j.fcr.2012.09.009

Watson-Hernández, F., Gómez-Calderón, N., & Da Silva, R. P. (2022). Oil palm yield estimation based on vegetation and humidity indices generated from satellite images and machine learning techniques. *AgriEngineering*, *4*(1), 279–291. https://doi.org/10.3390/agriengineering4010019

Yang, X., Hua, Z., Li, L., Huo, X., & Zhao, Z. (2024). Multi-source information fusion-driven corn yield prediction using the random forest from the perspective of agricultural and forestry economic management. *Scientific Reports*, *14*(1), 4052. https://doi.org/10.1038/s41598-024-54354-9