Ecological Informatics 81 (2024) 102595

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Predicting oil palm yield using a comprehensive agronomy dataset and 17 machine learning and deep learning models

Ehsan Jolous Jamshidi [a], Yusri Yusup [a,*], Chee Wooi Hooy [b], Mohamad Anuar Kamaruddin [a], Hasnuri Mat Hassan [c], Syahidah Akmal Muhammad [a], Helmi Zulhaidi Mohd Shafri [d], Kek Hoe Then [e], Mohd Shahkhirat Norizan [e], Choon Chek Tan [e]

[a] Environmental Technology, School of Industrial Technology, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia
[b] School of Management, School of Industrial Technology, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia
[c] School of Biological Sciences, Universiti Sains Malaysia, 11800, Pulau Pinang, Malaysia
[d] Department of Civil Engineering, Universiti Putra Malaysia, 43400, UPM-Serdang, Malaysia
[e] FGV R&D Sdn. Bhd., Malaysia

ARTICLE INFO

Keywords:
Oil palm
Yield prediction
Machine learning
Deep learning
Agriculture

ABSTRACT

The rising global demand for oil palm emphasizes the importance of accurate oil palm yield predictions. This predictive capability is critical for effective crop management, supply chain optimization, and sustainable farming practices. However, the oil palm sector faces challenges in yield projection, stressing a noteworthy gap in the application and evaluation of modern machine learning and deep learning technologies. Our study addressed this gap by systematically evaluating 17 machine and deep learning models in predicting oil palm yield, incorporating various agronomic parameters, e.g., soil composition, climatic conditions, plant age, and farming techniques. This holistic approach enhances the application of machine and deep learning in agriculture. Using the feature selection technique and a maximum depth of 32 and 1000 estimators, the Extra Trees Regressor exhibited positive performance, i.e., MSE = 860.36 and an R^2 = 0.65, and stands out among the 17 models evaluated. Our findings also showed that incorporating a comprehensive agronomic dataset is critical to accurate yield prediction. Hence, this model and approach have the potential to be a robust decision-making tool for agronomists and farmers in the oil palm industry, setting the stage for future innovations in sustainable agriculture practices.

1. Introduction

Oil palm (*Elaeis guineensis*) is globally recognized as the most productive oil crop, surpassing other oil crops like soybean, sunflower, and rapeseed. Specifically, oil palm yields 3.30 t per hectare, whereas soybean, sunflower, and rapeseed produce 0.46, 0.66, and 1.33 t, respectively (Basiron and Weng, 2004). In 2013, palm oil accounted for 44.6% of all oils, reflecting its significant market share (Bruno, 2017). Globally, palm oil is also the most widely consumed vegetable oil, constituting 35% of the world's vegetable oil consumption (Chong et al., 2017). The importance of oil palm is highlighted by its diverse uses, from culinary applications as cooking oil to its role in cosmetics, industrial lubricants, and biofuels (Koh and Wilcove, 2007). This dominance is evident from its production growth, increasing from a modest 1.8 Mt. in 1968 to 65.15 Mt. in 2018 (Rajakal et al., 2021), and the demand is rising by 7% annually, outpacing other vegetable oils at 4% (Bruno, 2017).

Oil palm is a major crop in Asia, particularly in Southeast Asian countries like Malaysia, Indonesia, and Thailand. These countries account for 85% of the global palm oil production (Iskandar et al., 2018). In these countries, oil palm plantations cover millions of hectares of land and contribute significantly to their economies. Oil palm is Asia's most widely produced vegetable oil, making up 37% of the total production of vegetable oils in 2013. The annual growth rate of oil palm production in Asia from 2004 to 2013 was 6.7% and the major contributors to global oil palm production are Indonesia and Malaysia, which reported 81.7% on average of the total global production during the same period (Oettli et al., 2018).

Malaysia has shifted its agricultural focus from rubber to oil palm cultivation. Data from the Malaysian Palm Oil Board (MPOB) reveals that in 2022, Malaysia planted oil palm across 5.67 million hectares.

* Corresponding author.
E-mail address: yusriy@usm.my (Y. Yusup).

https://doi.org/10.1016/j.ecoinf.2024.102595
Received 21 December 2023; Received in revised form 6 April 2024; Accepted 6 April 2024
Available online 17 April 2024

Recognizing the significant role of oil palm, extensive research has been conducted, as evidenced by studies from Melidawati et al. (2021), Tuerxun et al. (2020), and Ang et al. (2022), each contributing insights into oil palm cultivation and management. Predicting oil palm crop yields is crucial to effectively managing and optimizing the palm supply chain. It requires the incorporation of multiple datasets and is influenced by numerous variables, including climate patterns, soil quality, fertilizer use, and seed varieties (Xu et al., 2019). Applying Machine learning and deep learning models can accurately predict crop yields, aiding farmers in harvest planning and crop management decisions (Attri et al., 2023). Although these techniques have been implemented across various crops for environment-based yield forecasting, their deployment for predictive analysis of oil palm yields remains in its infancy (Rajakal et al., 2021).

Machine learning has become a powerful tool in the agricultural sector because it can identify patterns without being strictly tied to predefined models. Several recent studies in agriculture highlight the use of machine learning and deep learning for yield prediction. van Klompenburg et al. (2020) presented a systematic review of the application of machine learning in the yield prediction of various crops, focusing on algorithms like Neural Networks, Linear Regression, Random Forest, Support Vector Machine, and Gradient Boosting Tree, and utilizing features such as temperature, rainfall, soil type, crop information, and nutrients. Other works include researchers from the Shahhosseini et al. (2021) group, who used a hybrid method that combines simulation crop modeling and machine learning techniques to improve corn yield forecasts in Illinois, Indiana, and Iowa, USA. Gómez et al. (2019) highlighted the capability of remote sensing data and advanced machine learning models for precise potato yield predictions. Guo et al. (2021) focused on rice yield prediction using multiple linear regression (MLR) and three machine learning (ML) techniques, utilizing eleven data sets that integrated phenology, climate, and geography. Similarly, Cao et al. (2021) targeted rice yield prediction in China for the period 2000 to 2015 by leveraging the LASSO (statistical), random forest (machine learning), and LSTM (deep learning) methods. Their data sources encompass satellite vegetation indexes, meteorological data, and soil properties. These studies collectively demonstrated the potential of machine learning and deep learning techniques in agricultural yield prediction. They reveal that integrating diverse data sources like climate, soil properties, and remote sensing with advanced algorithms, such as Neural Networks, Random Forest, and LSTM can significantly enhance the accuracy of crop yield forecasts.

For oil palm, yield prediction exhibited unique challenges and methodologies that distinguish it from other crops because of specific factors, such as its perennial nature, canopy architecture, pollination mechanisms, and response to environmental stresses. Researchers have proposed a few techniques to forecast yield accurately. Some models can forecast yields up to nine months ahead. Awan and Sap (2006) developed a machine learning system to predict oil palm yield based on climate and plantation data. They suggested a weighted kernel k-means algorithm for unsupervised data partitioning that addresses noise, outliers, and autocorrelation in spatial data. The study demonstrated the algorithm's usefulness in analyzing the data and its utility in predicting oil palm yield. Keong and Keng (2012) developed a yield prediction model for oil palm plantations using agrometeorological factors; however, the model only explained 68% yield variability based on multiple linear regression, considering palm age and soil water capacity. Enhancements are suggested by considering terrain and soil traits because soil moisture at specific pre-harvest periods affects physiological stages.

Deep learning methods have shown some promise. Kartika et al. (2016) developed a three-layer neural network to forecast oil palm yield using weather data. The study found that model accuracy varied with the number of hidden layers, achieving the lowest error with five layers. Factors like rainfall and temperature were significant in yield prediction, while inaccurate data decreased accuracy. The model utilized a feed-forward neural network with backpropagation for minimum error predictions. Adjusting the hidden nodes optimized accuracy and validated against current yield data. Chapman et al. (2018) explored Bayesian networks for predicting oil palm yields using environmental and agronomic factors. Data from three Kalimantan estates informed models for fresh fruit bunch yield and other parameters. Ten-fold cross-validation tested accuracy, showing reliable predictions.

Hilal et al. (2020) explored the effectiveness of Artificial Neural Networks (ANNs) and Nonlinear Autoregressive with Exogenous Inputs (NARX) models in predicting the yield of oil palm fresh fruit bunches in Peninsular Malaysia. The study emphasized the significance of temperature and humidity on yield and critiqued traditional models like Multi Linear Regression (MLR) for their inability to account for weather variations. It advocated for neural networks as a superior method for predicting fresh fruit bunch yield, addressing the shortcomings of previous nonlinear modeling approaches that lacked accuracy and failed to include detailed data on soil properties and crop growth.

Ahmad Latif et al. (2021) study utilized machine learning for oil palm breeding, focusing on yield, fruit bunch characteristics, oil traits, and stress resistance using Support Vector Regression with a Radial Basis Function kernel for accurate predictions. Their research highlights the potential of machine learning to improve breeding efficiency but notes challenges such as scarce phenotype data, few agriculture-focused machine learning studies, and the lack of research in this area. Watson-Hernández et al. (2022) used Landsat satellite imagery to forecast annual oil palm production in Central Pacific, Costa Rica. They analyzed 103 productive units and distinct areas within a plantation used for production using Neural Network and Random Forest models. Neural Network excelled with datasets with variations of oil palms, while Random Forest performed best for the ones containing specific oil palms. Key predictors included vegetation indices linked to plant hydration. The study stresses the role of oil palm progeny variability in predictions.

Khan et al. (2022) analyzed oil palm yields in Malaysia using diverse datasets, improving data quality through preprocessing. They found tree-based models (Extra Tree and AdaBoost) most effective for yield prediction, although the specific impacts of key features on yields were unclear. The study highlighted several gaps: underuse of machine learning in oil palm yield prediction, limited research on using artificial neural networks, the inability of existing statistical models to capture nonlinear data dependencies, and the need for a modular prediction workflow to enhance understanding and prediction of oil palm yields using historical data. Ang et al. (2022) developed more accurate oil palm yield prediction models by integrating remote-sensed, field, and weather data, including vegetation indices and land surface temperatures from Landsat, and addressing errors. They identified key predictors such as rainfall, temperature, and vegetation indices, with models using these predictors showing improved accuracy. The Deep Neural Network (DNN) model was the most accurate, explaining 91% yield variation. Their research predicted oil palm yields using machine learning and deep learning, demonstrating that models like XGBoost and DNN surpassed traditional models (MLR, RF, SVR) in prediction accuracy. However, they noted slight overfitting in some models and limited their focus to oil palm without considering other crops or aspects.

Therefore, it can be noted that several research gaps still exist in oil palm yield prediction. Firstly, the existing body of research on applying machine learning or deep learning in oil palm yield remains limited due to the scarcity of data. Previous studies also overly relied on remote sensing data, which may not fully capture the conditions influencing yield. Moreover, machine learning and deep learning models were reported to be susceptible to overfitting; thus, the specific environmental impacts of these models remain ambiguous. Additionally, the nonlinear relationship between oil palm yield and weather data presents a significant challenge in predicting its yield. These previous findings suggest the necessity for a holistic approach in the yield prediction model selection and development. The oil palm industry faces significant challenges in enhancing yield predictions due to a shortage of comprehensive datasets and in-depth research employing machine and deep learning techniques. These challenges stem from the need to integrate diverse data types, including weather conditions, the genetic information of progeny, nutrient levels, soil characteristics, remote sensing data (e.g., vegetative indices), plant physiological factors, and varied agricultural practices such as irrigation status and fertilization strategies. While addressing the issue of model overfitting through feature selection and cross-validation methods, our study also delineates the specific impacts of different factors using a feature-importance approach.

This study addresses the abovementioned limitations through a comprehensive data analysis of 49 oil palm plots in Malaysia with detailed remote sensing and field data. Another critical aspect of our research involved comparing the efficacy of many linear versus nonlinear modeling techniques in predicting the inherently nonlinear trends of oil palm yields. This comparison was elucidated through the analysis of various models, categorized into linear (Bayesian Ridge Regressor, Elastic Net, Ridge Regression, Stochastic Gradient Descent Regressor, Lasso Regression) and nonlinear models (Extra Trees Regressor, LightGBM Regressor, CatBoost Regressor, Random Forest Regressor, Histogram-Based Gradient Boosting Regressor, Bagging Regressor, Deep Neural Network, XGBoost Regressor, Gradient Boosting Regressor, AdaBoost Regressor, Support Vector Regressor, Decision Tree Regressor).

Thus, this research aims to fulfill three primary objectives to advance yield predictions within the oil palm industry through the integration of machine and deep learning techniques: 1) to compile and analyze a comprehensive dataset that encompasses a wide array of variables affecting oil palm yields, such as weather conditions, progeny, soil types, nutrient levels, and agricultural practices, including irrigation and fertilization strategies, 2) to refine predictive modeling by addressing overfitting through sophisticated feature selection and cross-validation methods, while also employing a feature-importance approach to quantify the impact of various factors on yield predictions, and 3) to compare the effectiveness of linear and nonlinear modeling techniques in accurately forecasting the complex, nonlinear trends observed in oil palm yields.

2. Materials and methods

2.1. The study area

The study area for this research encompasses an experimental plantation in Pahang in Peninsular Malaysia, a region known for its contribution to the country's agricultural output, particularly in oil palm cultivation. The research is conducted across 49 distinct plots, each varying in intrinsic characteristics such as progeny-planting materials, soil type, and treatment methodologies. Fig. 1 displays the mapping specifics for the study area.

2.2. Datasets

2.2.1. Remote sensing data

We performed all remote sensing data processing on the Google Earth Engine (GEE) platform, specifically acquiring Landsat 7 (Santos et al., 2023) images from the "LANDSAT/LE07/C01/T1" collection, covering the period from early 2003 to the end of 2013. Utilizing the Python-based Earth Engine API, we implemented a sequence of operations to compute the Normalized Difference Vegetation Index (NDVI) and other vegetative indices that reflect different vegetation health and stress aspects. NDVI, a standard metric derived from the red (B3) and near-infrared (B4) bands, is an indicator of photosynthetic activity and biomass. In our processing script, NDVI was calculated through the normalized difference function, highlighting dense foliage and active growth areas.

Alongside NDVI, which can be affected by factors like climate change and humanoid activities (Han et al., 2022), our script computed additional indices such as the Simple Ratio (SR), Wide Dynamic Range Vegetation Index (WDRVI), Soil-Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI), Green NDVI (GNDVI), Green Chlorophyll Vegetation Index (GCVI), and the Enhanced Vegetation Index (EVI). NDVI and EVI measure vegetation greenness as indicators of healthy vegetation (Joshi et al., 2023). These indices incorporated band combinations to mitigate atmospheric effects, adjust for soil background, and enhance the sensitivity to canopy structure and chlorophyll content.

The data was acquired monthly for a given set of coordinates representing specific agricultural fields. Our script buffered each point to capture a 30-m radius around the targeted location, ensuring that the satellite data is representative of field conditions. The monthly imagery was then aggregated to median values to smooth out anomalies and capture a typical representation of the vegetation state for each month.

The comprehensive data frame created at the end of this process encapsulated longitudinal spectral data points for each location, providing a temporal snapshot of the vegetation's condition throughout the year. This granular and high-dimensional dataset extracted from

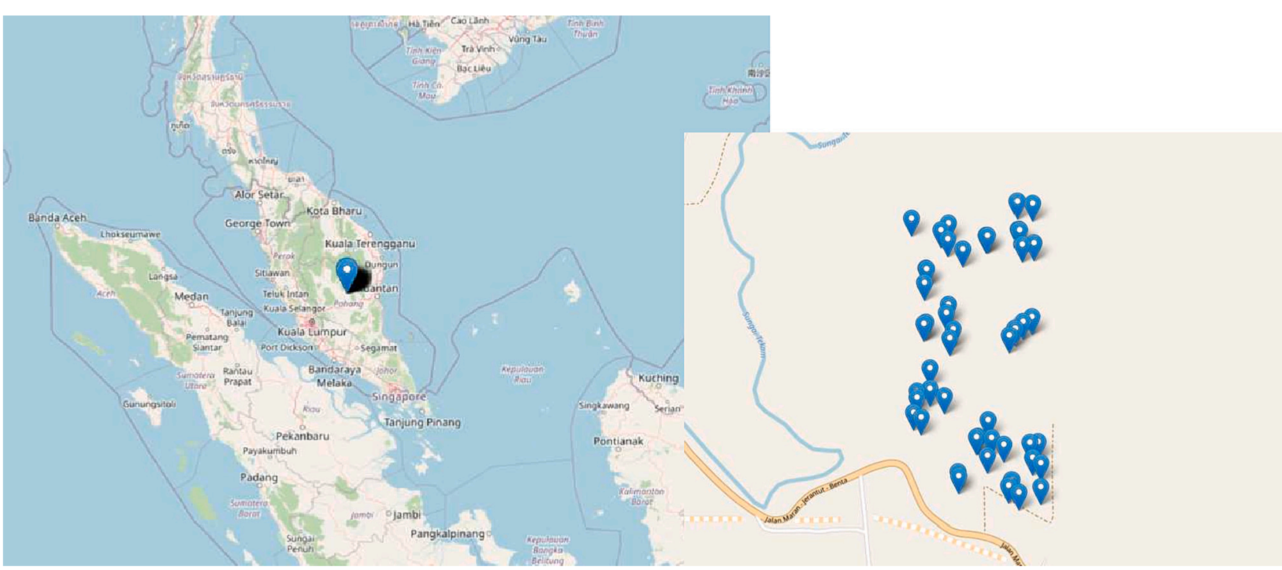

Fig. 1. Left: The location of the 49 experimental oil palm plots, Pahang, Peninsular Malaysia; Right: Each blue marker represents an experimental plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Landsat 7 imagery was needed in the training of our deep-learning models in discerning patterns regarding crop health and agricultural productivity.

2.2.2. Field data

Meteorological data is frequently used alongside various data types, especially for drought monitoring and agricultural production forecasting, due to its capacity to reflect changes in environmental conditions (Ferchichi et al., 2022). Our dataset included daily recordings of environmental conditions from January 2003 until December 2013. This includes rainfall in mm, maximum and minimum temperatures in °C, pan evaporation rates, and sunshine hours. Additionally, we have detailed records of soil types for each plot, which is essential for the understanding of the influence of edaphic conditions on crop growth and health.

The yearly data gathered provided plant growth and anatomy records through metrics, such as girth size (measured in m), frond production, canopy and rachis lengths (cm), leaflet number, and various other dimensions critical to understanding the vegetative state of the oil palm trees. Monthly nutrient data, including levels of phosphorus (P), potassium (K), boron (B), magnesium (Mg), and nitrogen (N) in the rachis and leaves, were also collected during the same period.

The dataset accounted for agricultural practices by including specifics such as the material (or progeny) of each plot, designated as S, L, N, D, Y, and A, reflecting a wide genetic variation. Information on the plot, irrigation treatment, age, and terrain type (undulating or terrace) was also recorded. Fertilizer applications from 2003 to 2013 were documented, detailing the year, round, type, month, and rate applied per palm.

Monthly production data, including the number of bunches and their corresponding weight, have been recorded. Yield calculations are derived from the bunch weight using the formula $\frac{\text{Bunch Weight} \times 136}{1000}$, which transformed bunch weight into a quantifiable yield metric.

The planting materials, categorized into six different types, present a diverse genetic canvas against Pahang's tropical climate. Despite each plot's unique characteristics, a uniform fertilizer program was applied, and climatic conditions were consistent across the board. This systematic approach ensures that yield variations can be ascribed with higher fidelity to specific experimental factors such as genetic variety and irrigation treatment. By amalgamating this extensive field data with remote sensing information, we created a robust dataset for the training of ML and DL models to predict oil palm yields annually.

2.3. Analysis workflow

This analysis utilized a structured methodology to predict oil palm yield (Fig. 2). The methodology encompasses several stages.

2.3.1. Data acquisition

The first stage was data acquisition, which involved gathering remote sensing data from Landsat 7 and various field data, including environmental factors, plant growth and anatomy, agricultural practices, and actual production outcomes, discussed in Section 2.2.

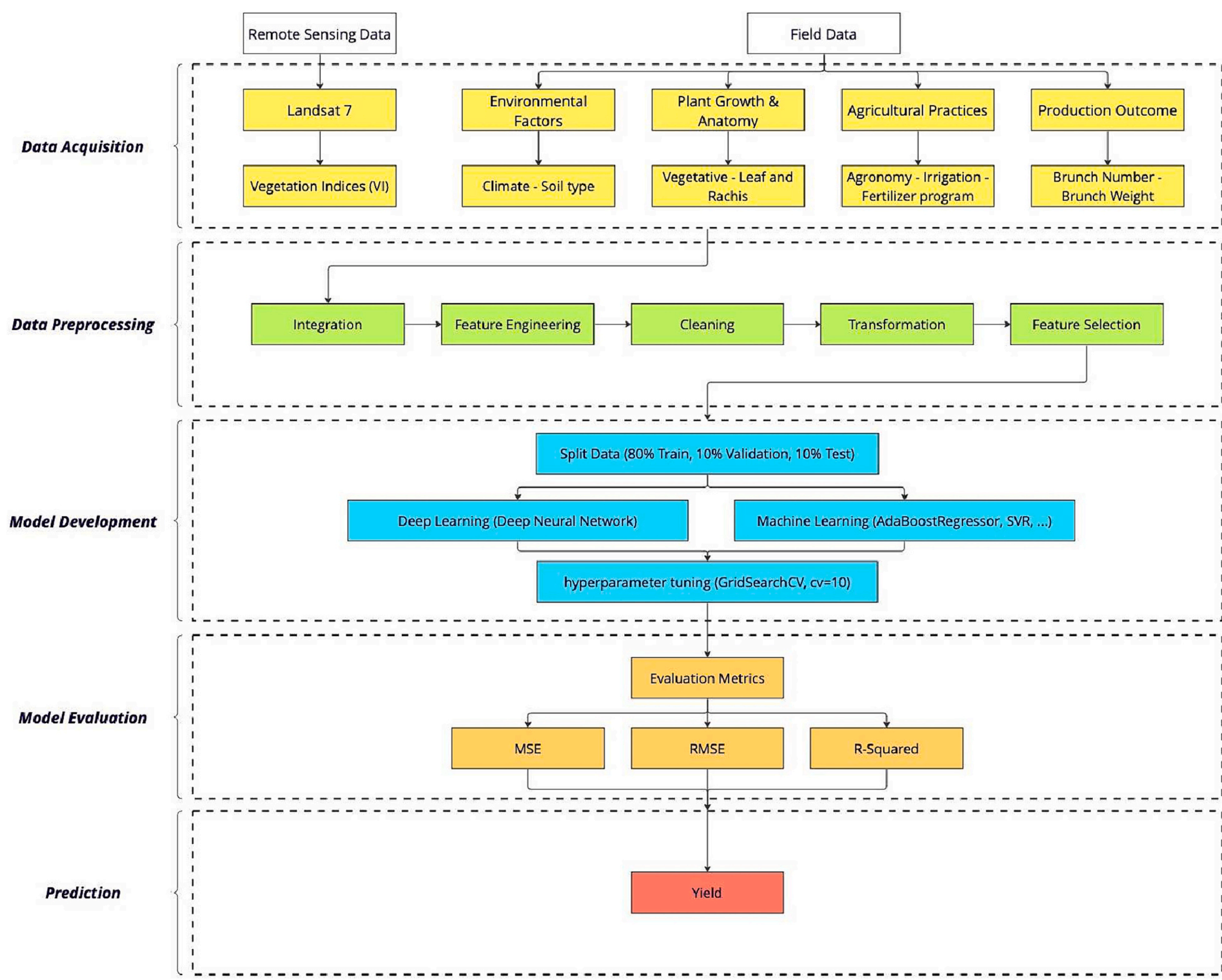

Fig. 2. The analysis workflow that shows the relationships, the data, and the methods employed in this study.

2.3.2. Preprocessing

In the second stage, data preprocessing integrated and refined the collected data through integration, feature engineering, cleaning, transformation, and feature selection. Various new features were engineered for the oil palm yield prediction model to capture the dynamic relationships between environmental factors and yield. These include indicators such as the "Chlorophyll Index," which serves as a proxy for the photosynthetic capacity of the plants and the temperature range, reflecting the diurnal thermal variations experienced by the crops. The "Soil Fertility Index" in the model captured the nutrient availability across different stages of the palm's growth cycle. Another feature, the "Crop Stress Index," incorporated climatic factors, such as temperature and rainfall, in conjunction with sunshine hours to quantify the environmental stress on the crop.

By engineering features that capture lagged weather effects, such as prior rainfall conditions, the model accounts for the temporal influence of weather on yield. These lagged features provided the model with information on the cumulative impact of weather patterns, acknowledging the fact that the effects of such variables are not always immediate but can have delayed consequences on crop development and yield.

We addressed missing values through the application of random forest imputation. This method estimated missing values using the random forest algorithm, which offers the advantage of handling the data's non-linearity and the interactions between variables. By leveraging the patterns found within the data, this imputation technique filled in missing values with estimates consistent with the rest of the dataset, maintaining the integrity of the data's structure. Moreover, during our feature selection process, we identified that the 'Bunch Number' column highly correlated with 'Bunch Weight.' Given that 'Bunch Weight' was a more significant predictor for oil palm yield and 'Bunch Number' would not be available for prediction in a practical scenario, we removed the 'Bunch Number' column from our dataset. This decision was based on the principle of parsimony, reducing multicollinearity and focusing on predictors available during prediction.

2.3.3. Transformation

Two fundamental techniques employed in the data transformation process are the MinMaxScaler and the LabelEncoder. The former is especially crucial for algorithms that are sensitive to the scale of the data, such as Support Vector Regressor (SVR) and neural networks. The LabelEncoder was applied to transform categorical fields into numerical values. Agricultural datasets often include categorical data, such as soil types or crop varieties, which must be encoded to numerical labels before being fed into machine learning algorithms.

2.3.4. Feature selection

Regression algorithms were explored to ascertain the most significant features influencing yield. Models ranged from ensemble methods like Extra Trees Regressor to various boosting and bagging techniques, including LightGBM Regressor and AdaBoost Regressor. The methodology was rigorous, employing Recursive Feature Elimination with Cross-Validation (RFECV) for feature ranking and selection, ensuring that the models utilized the most predictive variables. Techniques like Regularization were implicitly used to aid in identifying and discarding the features that contribute less to the predictive model, helping to prevent overfitting and improve generalization to new data.

Feature selection in agricultural data is crucial for multiple reasons, as highlighted by studies in S and R (2019) and Whitmire et al. (2021). It aids in pinpointing essential features that offer a thorough understanding of crop yield, such as days since sowing, Julian day, solar radiation, and rainfall. This process simplifies data collection and sheds light on the factors significantly impacting yield, like cumulative rainfall and days between harvest and sowing dates for crops. By focusing on the most relevant features, feature selection enhances model performance, enabling faster training, reduced complexity, improved interpretability, and accuracy while mitigating overfitting. This approach is more efficient, especially with vast datasets, by facilitating better outcomes with lesser computational efforts, underscoring its importance in improving agricultural productivity.

2.3.5. Model development

In the development stage, the preprocessed data was divided into training (80%), validation (10%), and test sets (10%) for model training and testing purposes (Shah et al., 2018), followed by the construction of predictive models using both deep learning algorithms (e.g., Deep Neural Networks) and traditional machine learning approaches (e.g., AdaBoost Regressor, Support Vector Regressor, etc.), with hyperparameter tuning performed via GridSearchCV with cross-validation (cv = 10). While some studies might use different proportions, such as a 70–30 split between training and testing/validation (Khan et al., 2022), the 80–10-10 split employed in this study ensured a more extensive training set, potentially leading to a better-generalized model. The test set, which the model has not previously seen during the training or validation phases indicates the model's performance in real-world scenarios (Shah et al., 2018).

The research integrates machine learning and deep learning models to enhance the accuracy of the forecasts. The models used can be clustered based on their approach, complexity, and underlying algorithms. In the machine learning category, ensemble models (AdaBoost Regressor, Extra Trees Regressor, CatBoost Regressor, Random Forest Regressor, and Bagging Regressor), linear models (Bayesian Ridge Regressor, Elastic Net, Lasso Regression, Ridge Regression, and Stochastic Gradient Descent Regressor), nonlinear models (Support Vector Regressor and Decision Tree Regressor), and gradient boosting models (XGBoost Regressor, LightGBM Regressor, Histogram-Based Gradient Boosting Regressor, and Gradient Boosting Regressor) were utilized. The deep learning model used was the deep neural network (Deep Neural Network). These algorithms were chosen because they are known for their ability to handle complex data, produce accurate predictions, and compare the accuracy of oil palm yield prediction methods.

Employing GridSearchCV ensured that each candidate combination of hyperparameters is assessed using cross-validation with a fixed number of folds, in this case, ten, as highlighted by Shahhosseini et al. (2021). This tenfold cross-validation provides a reliable estimate of the model's performance, balancing the trade-off between bias and variance and guarding against overfitting. The hyperparameter tuning process implemented in this study relied on a detailed and systematic search across a predefined grid of hyperparameter values. A unique parameter grid was defined for each model considered, targeting specific model characteristics such as the number of estimators in ensemble methods or the depth of trees in decision tree-based algorithms. This process ensured that the eventual choice of hyperparameters was not arbitrary but is backed by empirical evidence that aligns with Schratz et al. (2019), stating that optimal hyperparameter settings are critical to reducing bias in the model predictions.

2.4. Evaluation metrics

By applying statistical metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared R^2, the models' performances were assessed. The utilization of MSE and RMSE offers additional insights into the average magnitude of prediction errors, as evidenced by their widespread application in crop yield prediction (Khaki et al., 2020; Khaki and Wang, 2019; Mah and Nanyan, 2020). R^2 provides a normalized scale, making it particularly beneficial for comparing the goodness of fit across various models and datasets (Hilal et al., 2020). Eqs. (1) to (3) present the mathematical expressions for MSE, RMSE and R^2.

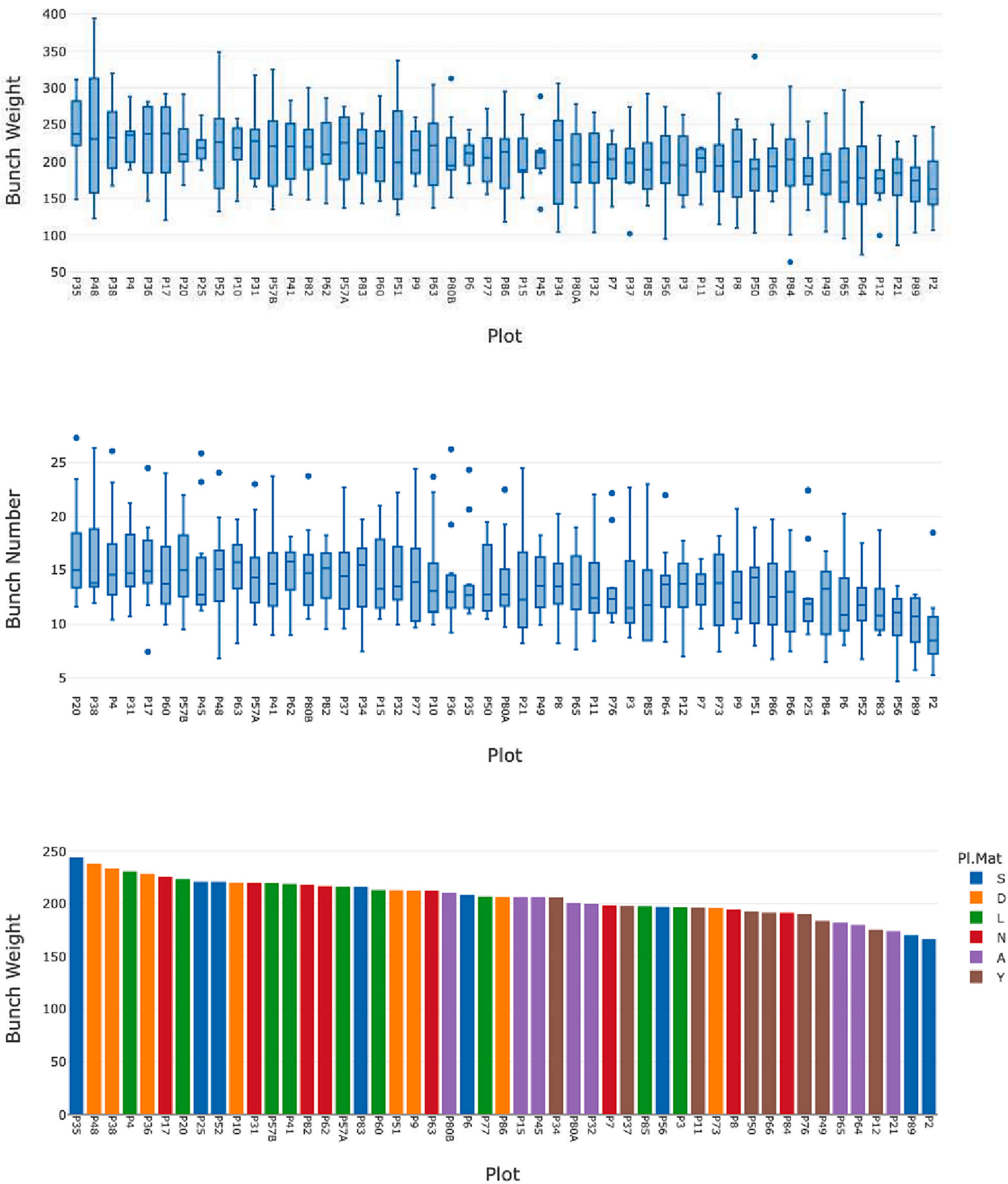

Fig. 3. The comprehensive analysis of oil palm yield: (Top) Distribution of oil palm bunch weight (BW) across 49 plots; (middle) Distribution of oil palm bunch number (BN) across 49; (bottom) average bunch weight per oil palm plot categorized by progenies, S, D, L, N, A, and Y.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_i^*\right)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^* - x_i\right)^2} \quad (2)$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}(x_i - \bar{x})\left(x_i^* - \bar{x}^*\right)}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2\sum_{i=1}^{N}\left(x_i^* - \bar{x}^*\right)^2}}\right)^2 \quad (3)$$

Here, $x_i, x_i^*, N, \bar{x}, \bar{x}^*$ represent the predicted value, actual value, sample size, mean predicted value, and mean actual value, respectively.

3. Results and discussion

This study presents a comprehensive analysis of an extensive agronomy dataset aimed at predicting oil palm yield using various machine and deep learning models. The dataset encompassed a diverse range of parameters collected from 538 samples of oil palm from 2003 to 2013. The age of these palms varied from 4 to 14 years, providing a substantial temporal scope for analysis. The data included critical factors, such as the chemical composition of the rachis and leaves, climatic conditions, physical plant attributes, vegetation indices, and direct measures of yield and biomass.

3.1. The variations among parameters and their association with oil palm yield

The dataset detailed the variability across 49 plots, each presenting unique characteristics. The analysis showed significant variations in parameters like rachis P (phosphorous) and K (potassium) levels, leaf N (nitrogen), P, K, Mg (magnesium), B (boron) levels, bunch weight, frond production, canopy and rachis measurements, leaf area, and dry weight per palm. These variations provided a rich source for developing and testing machine and deep learning models.

3.1.1. Chemical composition analysis

A detailed examination of the chemical composition of both rachis and leaves revealed a diverse nutritional status across the samples.

In the rachis, measured across 487 samples, phosphorus (P) levels averaged 0.099% (in dry matter) with a standard deviation of 0.033%, ranging from 0.035% to 0.276%. Potassium (K) levels in the rachis averaged 2.506% (in the dry matter) with a standard deviation of 0.926%, covering a range of 0.730% to 5.057%. The leaf analysis included N, P, K, Mg, and B measurements from 538 samples. The average nitrogen (N) level was 2.700% (on dry matter), with a standard deviation of 0.193% and a range from 2.020% to 3.280%. Phosphorus (P) in leaves averaged 0.168% (in dry matter) with a standard deviation of 0.011%, spanning from 0.145% to 0.213%. Potassium (K) levels were 1.048% (in dry matter) with a standard deviation of 0.186%, going from 0.395% to 2.195%. Magnesium (Mg) levels hovered at 0.210% (in dry matter), with a standard deviation of 0.058% and a range from 0.020% to 0.450%. Finally, the average boron (B) level was 15.651 mg/kg (in dry matter), with a standard deviation of 3.759 mg/kg and a range from 5.800 mg/kg to 34.000 mg/kg.

This chemical analysis indicated the varied nutritional conditions the oil palms experienced, which is crucial in understanding and predicting their yield.

3.1.2. Yield: Bunch weight (BW) and bunch number (BN)

A key part of the study involved the comparison of oil palm bunch weight across different plots, as illustrated in Fig. 3(top). This boxplot depicted the distribution of bunch weights for oil palms across several plots, denoted P1 through P89, where the weight was measured in kilograms. Each boxplot represented a different plot and displayed the minimum, first quartile (Q1), median, third quartile (Q3), and maximum bunch weight values. The data showed considerable variation both within and between plots. For instance, Plot P48 had the broadest range of bunch weights, suggesting a highly variable yield within this plot.

Conversely, Plot P4 displayed less variability with bunch weights ranging more narrowly. The median values across most plots hovered around the 200–240 kg range, suggesting a general level of consistency in median oil palm yields across these different plots. The average bunch weight also varied greatly, showing that different microclimates, soil fertility levels (Apori et al., 2020), or palm health might be influencing these variations.

Similarly, Fig. 3(middle) provided insights into the oil palm bunch number across different plots. The box plot visualized the distribution of oil palm bunch numbers across different plots, with each box representing the interquartile range (IQR) of the bunch number for a particular plot. The data for each plot, such as P20, P38, and P4, showed high variability regarding the average bunch number and the range of values. For example, plot P20 had the highest average bunch number at 16.71, while Plot P89 had the lowest at 10.14. The median values across the plots were relatively consistent, with many plots having medians in the 13 to 15 bunch range. This indicated a general level of consistency in median oil palm yields across these different plots. The IQR variation suggested production variability within the plot, while outliers in many plots indicated significant deviations from the typical production numbers.

When comparing the two plots in Figs. 3(top) and 3(middle), there was noticeable variability in weight and number of bunches across the different plots. This suggested that varying conditions across the plots could be influencing these measures.

Interestingly, no direct correlation was visible between the weight and number of bunches. For example, P35, which had the highest average bunch weight, did not have the highest average number of bunches, which P20 held. This indicated that factors that increase weight do not necessarily increase the number of bunches and vice versa. Some plots showed consistency in their position across both measures, such as Plot P2, which was at the lower end in the average bunch weight and number, indicating consistent underperformance in both aspects.

3.1.3. Yield and progeny

We further explored the average bunch weight (BW) per plot by progeny, as shown in Fig. 3(bottom). This bar graph categorized the average BW for different oil palm plots by progeny, denoted by single letters representing different genetic varieties or crosses within the oil palm species. The graph highlighted variations in BW among different plots for identical progeny, indicating that environmental factors or differences in agronomic practices could significantly influence yield. For instance, the S progeny, particularly in plot P35, showed the highest average bunch weight, suggesting it might have a higher yield potential than other progenies. Progeny D consistently showed high bunch weights across multiple plots, indicating a possible robustness in this progeny's performance across different environments. Conversely, progenies represented by letters A and Y had the lowest average bunch weights across most plots, potentially indicating a lower genetic potential for yield or higher sensitivity to the environmental conditions of the tested plots (Sparnaaij, 1959).

3.1.4. Yield and its simultaneous association with environmental parameters

The heatmap analysis in Fig. 4 offers a visual summary of the average bunch weight across various plots, showcasing the influence of different agricultural treatments, terrains, progeny, and soil types. The plots marked with 'IR,' which denotes irrigation, generally showed higher average yields than 'NIR' without irrigation, suggesting that irrigation

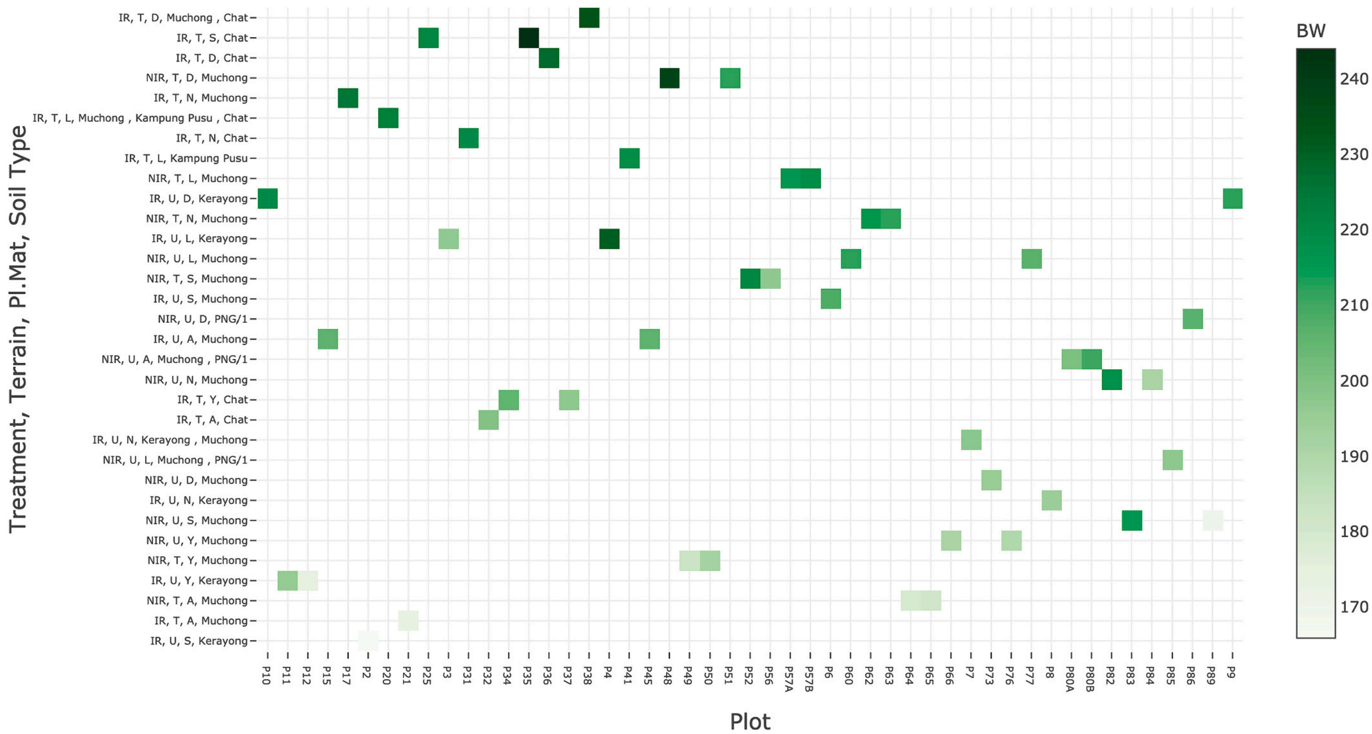

Fig. 4. Heatmap of average oil palm yield by plot, incorporating factors such as irrigation, terrain, parent material, and soil type.

positively affected the crops' bunch weight.

Terraced terrains (T) had a more consistent performance than undulating terrains (U), potentially due to improved water retention and reduced soil erosion associated with terraced agriculture. The diversity in soil types showed varied results, indicating that soil composition significantly impacted yield. Some soil types, like 'Muchong,' 'Kampung Pusu,' and 'Chat,' performed better, possibly due to better nutrient content or structure conducive to growth. The different letters representing the genetic lines also indicated variations in the genetic makeup of the crops that responded differently to environmental factors.

The correlation of BW with other numeric columns for each plot, as illustrated in Fig. 5, revealed the relationship between BW and various agronomic variables across different plots. The age of the palms showed a moderate to strong positive correlation with BW across most plots, suggesting that as palms mature, their potential to yield higher bunch weights increases. The correlation between nutrient content and BW varied considerably among the plots, indicating that the optimal nutrient levels for maximizing BW might vary significantly. Environmental conditions generally showed a negative correlation between higher temperatures and BW, suggesting extreme heat could harm the yield (Fleiss et al., 2022). Leaf metrics showed a strong positive correlation with BW in several plots, indicating that a larger photosynthetic area contributes to higher yield potential.

Remote sensing indices showed a strong positive correlation with BW in many plots, suggesting that these indices could be reliable predictors of yield and helpful in monitoring and managing the plantation (Balasundram et al., 2013). Frond production, canopy and rachis metrics, and water-related variables also showed varied correlations with BW. The negative correlations with the Crop Stress Index across most plots were consistent with the expectation that stress adversely affects yield (Neto et al., 2021). The Soil Fertility Index also had a mixed correlation with BW, reinforcing the idea that soil conditions need to be tailored to the specific needs of each plot for optimal yields.

The analysis of our study highlighted the complexity and variability inherent in predicting oil palm yield—the diverse range of factors influencing yield, from genetic makeup to environmental conditions.

3.2. *Machine learning models and agronomic parameters evaluation*

The approach bridged the gap between the detailed analysis of the oil palm dataset and the evaluation of machine learning models. It demonstrated the transition from empirical findings to predictive modeling. The comprehensive dataset, characterized by its extensive range of parameters, provided a robust foundation for applying advanced machine-learning techniques. This shift is not only a change in methodology but also a distinction in focus, i.e., from understanding the present and past patterns in oil palm yields to forecasting future outcomes.

3.2.1. *Comparison of 17 machine learning and one deep learning models*

The predictive models evaluated in this part of the study, such as the Extra Trees Regressor, Light-GBM Regressor, and CatBoost Regressor, were selected based on their ability to handle the complexity of the dataset. The performance of the models, assessed through validation mean square errors (MSE) and R^2 reflected their capacity to capture and predict the subtle interactions among agronomic factors.

The comparison of different models in Table 1, each with its strengths and limitations, features the intrinsic nature of modeling agricultural phenomena, where factors like soil composition, climatic conditions, and genetic variances are intertwined. Thus, this study provides a comprehensive understanding of oil palm yield factors and paves the way for advanced, data-driven approaches to agricultural forecasting.

Our results show that the Extra Trees Regressor emerged as the most accurate model, with an MSE of 860.36 and an R^2 of 0.65. Its best parameters included a maximum depth of 32 and 1000 estimators. The model employs a detailed approach to analyze oil palm yield determinants, focusing on key features such as palm age, Borate levels, and canopy length, alongside soil and climate indicators. Age is the most vital factor, closely linked to the production capacity of the palms. Borate levels reflected essential soil micronutrients, while canopy length indicated the health and vigor of the palms, affecting photosynthesis and yield. The chlorophyll and crop stress indexes also played significant

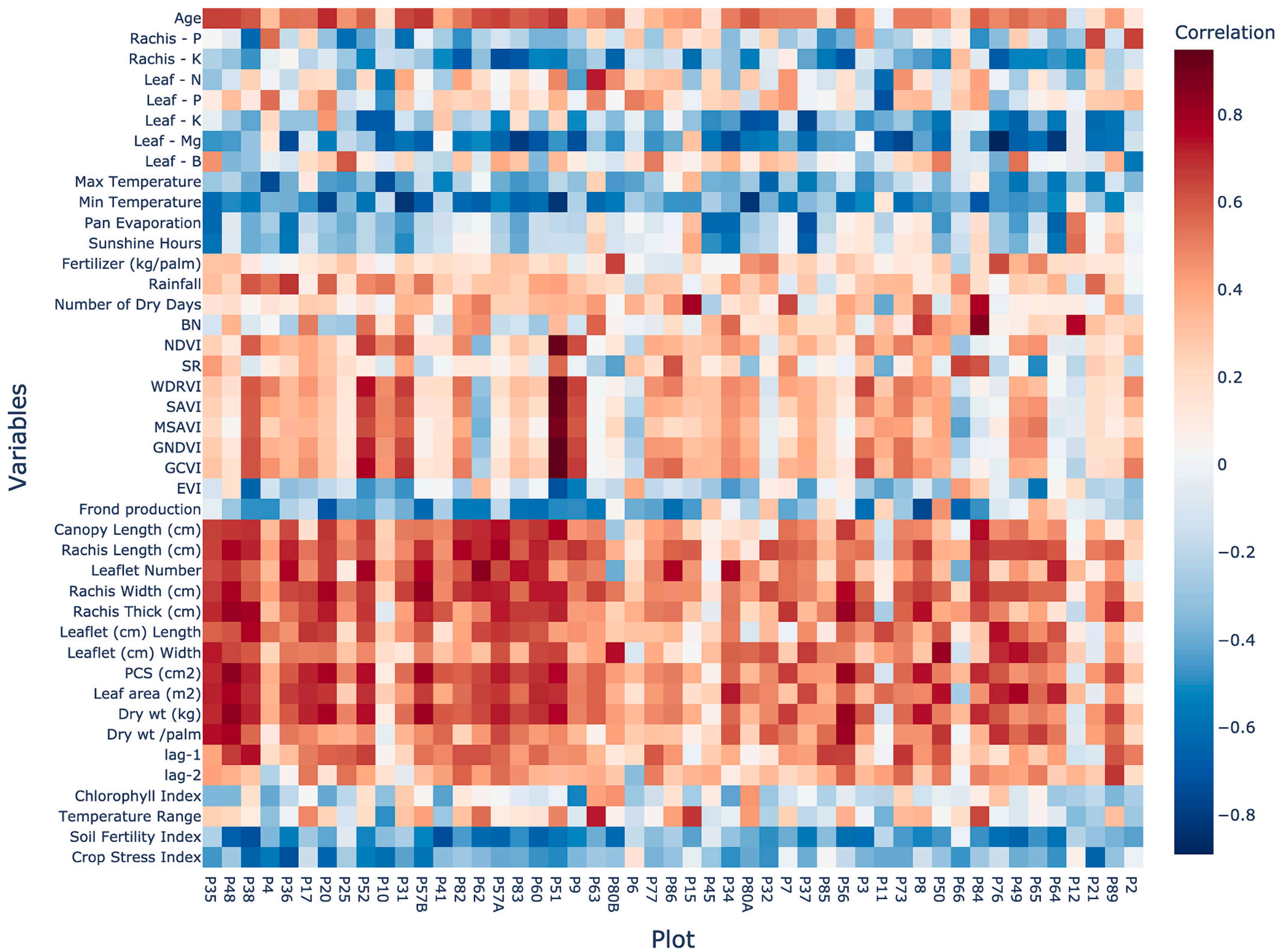

Fig. 5. Correlation heatmap of bunch weight and agronomic variables in oil palm plots.

Table 1
Performance comparison of regression models.

Model Name	MSE	RMSE	R^2
Extra Trees Regressor	860.36	29.33	0.65
LightGBM Regressor	891.03	29.85	0.64
CatBoost Regressor	991.96	31.49	0.60
Random Forest Regressor	1001.27	31.64	0.60
Histogram-Based Gradient Boosting Regressor	1016.18	31.87	0.59
Bagging Regressor	1025.74	32.02	0.59
Deep Neural Network	1082.91	32.90	0.56
XGBoost Regressor	1132.50	33.65	0.54
Gradient Boosting Regressor	1211.98	34.81	0.51
AdaBoost Regressor	1253.70	35.40	0.50
Bayesian Ridge Regressor	1254.74	35.42	0.49
Support Vector Regressor	1256.88	35.45	0.49
Elastic Net	1259.70	35.49	0.49
Ridge Regression	1260.54	35.50	0.49
Stochastic Gradient Descent Regressor	1278.36	35.75	0.49
Lasso Regression	1320.39	36.33	0.47
Decision Tree Regressor	1610.90	40.13	0.35

roles in explaining crop productivity and stress. These features collectively contribute to a comprehensive understanding of the factors affecting oil palm yield.

After establishing the Extra Trees Regressor as the top-performing model based on the lowest validation MSE and a robust R^2 value, further analyses of its predictive accuracy were conducted through a scatter plot of the actual versus predicted yields. This visual representation, as depicted in Fig. 6 in the left panel, shows the distribution of predicted values when plotted against the actual data, with the ideal prediction line in red serving as a benchmark for perfect accuracy. The proximity of the data points to this line reflects the precision of the model, and the scatter plot reveals that the Extra Trees Regressor maintains a consistent degree of accuracy across the range of values, with few outliers. It also excelled due to its method of randomly splitting data into numerous subgroups and making predictions from each, effectively handling data variability. This approach surpasses models like LightGBM Regressor by avoiding overfitting, demonstrating stability, and adaptively improving. Key parameters appreciably enhanced performance by optimizing data splitting and prediction accuracy, contributing to its robustness against overfitting and ensuring consistent improvement.

Further examination of the model's residuals, the differences between the actual yields and those predicted by the model, provides additional insight into its performance. The residuals histogram, illustrated in Fig. 6, right panel, exhibits a distribution with its peak close to zero, indicating that most of the model's predictions are accurate within a small margin of error. The symmetry of the distribution shows that the model does not exhibit systematic overestimation or underestimation across the dataset. These plots highlight the model's capacity to effectively generalize from the training data to the validation set, thereby reinforcing the validity of the Extra Trees Regressor for yield prediction tasks in oil palm agronomy.

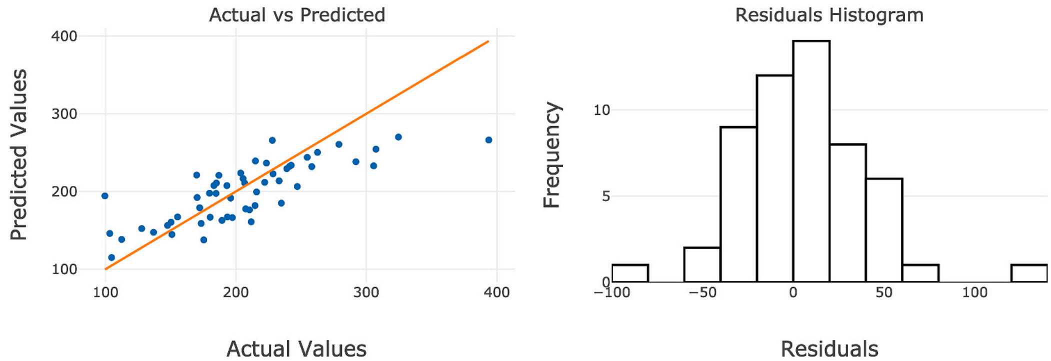

Fig. 6. Left: Scatter plot of the Extra Trees Regressor model showing the correlation between the actual and predicted oil palm yields. The red line indicates the line of perfect prediction. Right: Histogram of the residuals from the Extra Trees Regressor model, indicating the frequency of prediction errors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The LightGBM Regressor followed the Extra Trees Regressor in terms of performance, with an MSE of 891.03 and an R^2 of 0.64. Its configuration involved a learning rate of 0.05 and a maximum depth of 5, along with 200 estimators. The model preferred features that span from canopy length and chlorophyll index to soil types and climate factors, demonstrating a balanced consideration of plant-specific and environmental variables.

The CatBoost Regressor also exhibited commendable performance, with an MSE of 991.96 and an R^2 of 0.60. Its depth of 8 and iteration count of 10, coupled with a diverse feature set encompassing canopy length, crop stress index, and various soil and climatic parameters, played a critical role in its predictive capability.

In stark contrast, the Decision Tree Regressor recorded negligible efficacy, with an MSE of 1610.90 and an R^2 of 0.35. This model's limited depth of 3 and a minimalistic approach to sample splitting and leaf parameters contributed to its underwhelming performance.

3.2.2. *Possible reasons for the varying performances of the models*

Examining the other models in the study offers a broad view of their relative performances. Models like the Random Forest Regressor, Histogram-Based Gradient Boosting Regressor, and Bagging Regressor exhibited moderate performances. Their MSE values ranged from 1001.27 to 1025.74, and R^2 values hovered around the 0.6 mark. On the other end of the spectrum, models such as the Deep Neural Network, XGBoost Regressor, and Gradient Boosting Regressor showed lower effectiveness, with MSE values exceeding 1100 and R^2 values dipping below 0.55. These variations emphasize the differences in how each model architecture interacts with the dataset and the complexity of oil palm yield prediction.

It becomes evident that tree-based models, like the Extra Trees Regressor (Geurts et al., 2006) and Random Forest Regressor (Verma and Dong, 2016), excelled in capturing nonlinear relationships and interactions among features due to their architectures. Their advantage is their ability to handle many features without substantial feature engineering.

Boosting models like the LightGBM Regressor and XGBoost Regressor, known for their sequential correction of the errors of their predecessors, have shown a strong ability to improve accuracy, albeit at the risk of overfitting (Chen and Guestrin, 2016; Ke et al., 2017).

Neural networks represented by the Deep Neural Network did not perform optimally, possibly due to the limitations in the dataset size or feature representation (Davila Delgado and Oyedele, 2021). This is surprising given that neural networks are robust in handling complex data.

The choice and tuning of parameters played a critical role in the performance of these models (Yang and Shami, 2020). Parameters like 'max_depth' in tree-based models and 'learning_rate' in boosting models impacted the ability of the model to learn the data without overfitting. Similarly, the number of estimators or iterations controlled the extent of learning, making it a crucial parameter for model optimization.

The selection of features for each model also sheds light on the complexity of predicting oil palm yield. The importance of features related to soil type, climate conditions, and plant-specific metrics like canopy length and chlorophyll index indicates the multidimensional nature of the factors influencing oil palm yield. Notably, while some models benefited from a more extensive set of features, others achieved comparable performance with a more streamlined feature set, suggesting that the quality of features can be as critical as their quantity (Gudivada et al., 2017).

In the context of predictive model performance, it is imperative to consider the impact of the limited number of observations for each plot. A constrained dataset size can significantly influence model accuracy, as it restricts the model's ability to learn and generalize from the data effectively. This limitation is particularly pertinent in complex agricultural systems like oil palm yield prediction, where the diversity and interplay of numerous variables demand a robust dataset for optimal model training and validation. Therefore, the relatively lower performance of some models in this study could be partly attributed to insufficient observations, which hampers the models' capability to capture and predict the intricate patterns and interactions inherent in the agricultural data. This factor stresses the need for more extensive and comprehensive datasets to enhance the predictive accuracy of machine learning models in agricultural forecasting.

In summary, this study analyzes various machine learning and deep learning models applied to predict oil palm yield. The standout performance of models like the Extra Trees Regressor and LightGBM Regressor, as opposed to the less effective Decision Tree Regressor, highlights the significance of model architecture, parameter optimization, and feature selection in tackling complex agricultural prediction tasks.

4. Conclusions

We comprehensively analyzed 17 machine learning and deep learning models for oil palm yield prediction by utilizing a robust agronomy dataset encompassing a multitude of parameters collected from 538 oil palm samples. This extensive dataset served as the foundation for exploring the predictive accuracies of different modeling approaches.

Central to our findings was the performance of the Extra Trees Regressor model. With a maximum depth of 32 and 1000 estimators, this model emerged as the most effective tool. Through scatter plots of actual versus predicted yields, confirmation of the model's effectiveness was obtained with an intuitive understanding of its predictive behavior in real-world scenarios. It achieved a Mean Squared Error (MSE) of 860.36 and an R^2 value of 0.65, using agronomic factors, from Age and Borate to soil and climate.

Indicators were critical in capturing the determinants of oil palm yield. Models like the Random Forest Regressor, Histogram-Based Gradient Boosting Regressor, and Bagging Regressor exhibited moderate performance levels, with MSE values ranging from 1001.27 to 1025.74 and R^2 values around the 0.6 mark. Conversely, the Deep Neural Network, XGBoost Regressor, and Gradient Boosting Regressor models demonstrated lower effectiveness, with MSE values exceeding 1100 and R^2 values dipping below 0.55.

The study analyzes the effects of soil composition, climate, plant age, and agricultural practices on oil palm yield, identifying significant impacts from soil nutrient variability, climatic conditions, and plant age, but also highlights the critical role of feature selection in yield prediction models. It demonstrated how factors like soil type, climate conditions, and plant-specific metrics such as canopy length and chlorophyll index are essential, illustrating the multidimensional nature of influences on oil palm yield. While some models use more features for improved performance, others achieve comparable results with streamlined feature selection, underlining the importance of managing soil nutrients, adapting to climate variability, understanding age-related yield patterns, and optimizing agronomic practices for enhanced productivity.

Our study offers valuable insights into the potential and challenges of applying machine learning and deep learning models to oil palm yield prediction. The Extra Trees Regressor model showcases the effectiveness of machine learning techniques in dealing with complex, real-world problems. However, it also brings to the fore the importance of careful feature selection and the need for a tempered understanding of the diverse factors that influence agricultural yields.

This research underscores the significant impact of machine learning (ML) and deep learning (DL) models on the oil palm sector, offering a path to precise yield predictions and sustainable crop management. It highlights the utility of these models in enhancing agricultural practices by considering soil composition, climate variations, and farming techniques, thus aligning production with market demands and optimizing the supply chain. Agronomists are encouraged to adopt a comprehensive approach leveraging ML and DL for research and development. At the same time, farmers can benefit from data-driven insights, mainly through the Extra Trees Regressor model, to improve planning and resource management. Additionally, the findings serve as a basis for stakeholders, including policymakers and the agribusiness sector, to support ML integration in agriculture, fostering operational efficiency, sustainability, and market stability through improved supply chain management and risk mitigation.

4.1. Limitation and future studies

The research encountered specific limitations that warrant attention for future studies. A notable limitation stems from the diversity in data elements, where the dataset encompasses a wide range of tree families, soil types, and irrigation systems. This variety introduces high variability, challenging the models' ability to discern clear patterns, which could lead to lower predictive accuracy as indicated by reduced R2 values. Additionally, the dataset's size is insufficient, comprising only 49 plots observed over 11 years, yielding less than 550 data points. Such a limited dataset is inadequate for high-dimensional data analysis, hindering the practical application of robust models like deep forests or gradient-boosting machines.

Furthermore, the regional span of the study is small, and the results may only apply to that particular area. This geographic limitation is constraining and suggests a gradual expansion to ensure broader applicability of the findings. To address these issues, future research could focus on consolidating similar data elements to reduce variability and exploring methods to generate synthetic data and incorporate additional relevant data. These approaches aim to enrich the dataset, potentially enhancing model training and performance by providing a more comprehensive foundation for identifying underlying patterns.

Ethical approval

Not Applicable.

Availability of supporting data

Not Applicable.

Funding

Ehsan Jolous Jamshidi and Yusri Yusup from Universiti Sains Malaysia were funded by the Ministry of Higher Education, Malaysia. Hooy Chee Wooi, Mohamad Anuar Kamaruddin, Hasnuri Mat Hassan, and Syahidah Akmal Muhammad, also from Universiti Sains Malaysia, received no funding. Helmi Zulhaidi Mohd Shafri from the Department of Civil Engineering, Universiti Putra Malaysia, also received no funding. Then Kek Hoe, Mohd Shahkhirat Norizan, and Tan Choon Chek from FGV R&D Sdn. Bhd. also received no funding for this study.

Authors' contributions

E.J.J. (Ehsan Jolous Jamshidi): Conceived the study, led the data analysis, and wrote the manuscript.

Y.Y. (Yusri Yusup): Provided expertise on agricultural datasets, contributed to the study design, edited the manuscript, and served as the corresponding author.

H.C.W. (Hooy Chee Wooi): Focused on the economic and environmental aspects of the research, contributed to the literature review and reviewed the manuscript.

M.A.K. (Mohamad Anuar Kamaruddin): Assisted in data preprocessing, especially data cleaning and normalization, and edited the manuscript.

H.M.H. (Hasnuri Mat Hassan): Contributed to the biological aspects of crop prediction, particularly in feature extraction and selection, and reviewed the manuscript.

S.A.M. (Syahidah Akmal Muhammad): Contributed to the biological aspects of crop prediction, particularly in feature extraction and selection, and reviewed the manuscript.

H.Z.M.S. (Helmi Zulhaidi Mohd Shafri): Provided technical expertise in remote sensing and GIS applications in agriculture.

T.K.H. (Then Kek Hoe): Involved in the case study and provided the dataset.

M.S.N. (Mohd Shahkhirat Norizan): Prepared and provided the agronomy data.

T.C.C. (Tan Choon Chek): Provided the dataset, guided the data analysis and reviewed the manuscript. All authors reviewed and approved the final manuscript.

CRediT authorship contribution statement

Ehsan Jolous Jamshidi: Writing – original draft, Formal analysis, Conceptualization. **Yusri Yusup:** Writing – review & editing, Project administration, Methodology. **Chee Wooi Hooy:** Writing – review & editing, Validation, Investigation. **Mohamad Anuar Kamaruddin:** Writing – original draft, Data curation. **Hasnuri Mat Hassan:** Investigation. **Syahidah Akmal Muhammad:** Validation, Investigation. **Helmi Zulhaidi Mohd Shafri:** Resources. **Kek Hoe Then:** Writing – review & editing, Resources, Methodology. **Mohd Shahkhirat Norizan:** Data curation. **Choon Chek Tan:** Writing – review & editing, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

The authors do not have permission to share data.

Acknowledgments

We acknowledge the Ministry of Education Malaysia for awarding us the Malaysian Research University Network Long-Term Research Grant Scheme (MRUN-LRGS) with grant number 203.PTEKIND.6777006 that made this research possible.

References

Ahmad Latif, N., Mohd Nain, F.N., Ahamed Hassain Malim, N.H., Abdullah, R., Abdul Rahim, M.F., Mohamad, M.N., Mohamad Fauzi, N.S., 2021. Predicting heritability of oil palm breeding using phenotypic traits and machine learning. Sustainability 13 (22).

Ang, Y., Shafri, H.Z.M., Lee, Y.P., Bakar, S.A., Abidin, H., Mohd Junaidi, M.U.U., 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 Sci. Inf. 15 (4), 2349–2367.

Apori, S.O., Sadick, A., Hanyabui, E., Musah, M., Marius Flarian, M., Acheampong, M., 2020. Evaluation of soil fertility status in oil palm plantations in the western region of Ghana. AIMS Agricult. Food 5, 938–949.

Attri, I., Awasthi, L.K., Sharma, T.P., Rathee, P., 2023. A review of deep learning techniques used in agriculture. Eco. Inform. 77, 102217.

Awan, A.M., Sap, M.N.M., 2006. A framework for predicting oil-palm yield from climate data. In: In Proceedings of the Postgraduate Annual Research Seminar, pp. 360–366.

Balasundram, S., Memarian, H., Khosla, R., 2013. Estimating oil palm yields using vegetation indices derived from quickbird. Life Sci. J. 10, 851–860.

Basiron, Y., Weng, C.K., 2004. The oil palm and its sustainability. J. Oil Palm Res. 16 (1).

Bruno, L.C., 2017. Palm oil plantation productivity during the establishment of the malaysian refinery sector, 1970–1990. Econ. Hist. Develop. Reg. 32 (3), 221–269.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., Xie, J., 2021. Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agric. For. Meteorol. 297, 108275.

Chapman, R., Cook, S., Donough, C., Lim, Y.L., Vun Vui Ho, P., Lo, K.W., Oberthür, T., 2018. Using bayesian networks to predict future yield functions with data from commercial oil palm plantations: A proof of concept analysis. Comput. Electron. Agric. 151, 338–348.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable Tree Boosting System. *CoRR*, abs/1603.02754.

Chong, K.L., Kanniah, K.D., Pohl, C., Tan, K.P., 2017. A review of remote sensing applications for oil palm studies. Geo-spat. Inf. Sci. 20 (2), 184–200.

Davila Delgado, J.M., Oyedele, L., 2021. Deep learning with small datasets: using autoencoders to address limited datasets in construction management. Appl. Soft Comput. 112, 107836.

Ferchichi, A., Abbes, A.B., Barra, V., Farah, I.R., 2022. Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. Eco. Inform. 68, 101552.

Fleiss, S., McClean, C.J., King, H., Hill, J.K., 2022. Limited impacts of climatic conditions on commercial oil palm yields in malaysian plantations. CABI Agric. Biosci. 3 (1), 59.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sens. 11 (15).

Gudivada, V., Apon, A., Ding, J., 2017. Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int. J. Adv. Software 10, 1–20.

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Robin Bryant, C., Senthilnath, J., 2021. Integrated phenology and climate in rice yields prediction using machine learning methods. Ecol. Indic. 120, 106935.

Han, F., Fu, G., Yu, C., Wang, S., 2022. Modeling nutrition quality and storage of forage using climate data and normalized-difference vegetation index in alpine grasslands. Remote Sens. 14 (14), 3410.

Hilal, Y.Y., Yahya, A., Ismail, W.I.W., Asha'ari, Z.H., 2020. Neural networks method in predicting oil palm ffb yields for the peninsular states of Malaysia. J. Oil Palm 33, 400–412.

Iskandar, M.J., Baharum, A., Anuar, F.H., Othaman, R., 2018. Palm oil industry in south east asia and the effluent treatment technology—a review. Environ. Technol. Innov. 9, 169–185.

Joshi, A., Pradhan, B., Chakraborty, S., Behera, M.D., 2023. Winter wheat yield prediction in the conterminous United States using solar-induced chlorophyll fluorescence data and xgboost and random forest algorithm. Eco. Inform. 77, 102194.

Kartika, N.D., Astika, I.W., Santosa, E., 2016. Oil palm yield forecasting based on weather variables using artificial neural network. Indonesian J. Elect. Eng. Comp. Sci. 3 (3), 626.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc.

Keong, Y.K., Keng, W.M., 2012. Statistical modeling of weather-based yield forecasting for young mature oil palm. In: APCBEE Procedia, 4:58–65. 2nd International Conference on Asia Agriculture and Animal (ICAAA 2012).

Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks. Front. Plant Sci. 10.

Khaki, S., Wang, L., Archontoulis, S.V., 2020. A cnn-rnn framework for crop yield prediction. Front. Plant Sci. 10.

Khan, N., Kamaruddin, M.A., Ullah Sheikh, U., Zawawi, M.H., Yusup, Y., Bakht, M.P., Mohamed Noor, N., 2022. 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).

Koh, L.P., Wilcove, D.S., 2007. Cashing in palm oil for conservation. Nature 448 (7157), 993–994.

Mah, P.J.W., Nanyan, N.N., 2020. A Comparative Study between Univariate and Bivariate Time Series Models for Crude Palm Oil Industry in Peninsular Malaysia.

Melidawati, Cherie, D., Fahmy, K., Makky, M., 2021. Non destructive evaluation quality of oil palm fresh fruit bunch (ffb) (elaeis guineensis jack) based on optical properties using artificial neural network (ann). IOP Conf. Series: Earth Environ. Sci. 644 (1), 012032.

Neto, J.C.R., Vieira, L.R., de Aquino Ribeiro, J.A., de Sousa, C.A.F., Júnior, M.T.S., Abdelnur, P.V., 2021. Metabolic effect of drought stress on the leaves of young oil palm (elaeis guineensis) plants using uhplc–ms and multivariate analysis. Sci. Rep. 11 (1), 18271.

Oettli, P., Behera, S.K., Yamagata, T., 2018. Climate based predictability of oil palm tree yield in Malaysia. Sci. Rep. 8 (1).

Rajakal, J.P., Andiappan, V., Wan, Y.K., 2021. Mathematical approach to forecast oil palm plantation yield under climate change uncertainties. Chem. Eng. Trans. 83, 115–120.

S, M.G.P., R, B., 2019. Selection of important features for optimizing crop yield prediction. Int. J. Agricult. Environ. Inform. Syst. 10 (3), 54–71.

Santos, J.S., de Mendonça, A.R., Gonçalves, F.G., da Silva, G.F., de Almeida, A.Q., de Pádua Chaves e Carvalho, S., Silva, J.P.M., Carvalho, R.C., da Silva, E.F., Aguiar, M. O., 2023. Predicting eucalyptus plantation growth and yield using landsat imagery in Minas Gerais, Brazil. Eco. Inform. 75, 102120.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120.

Shah, D., Campbell, W., Zulkernine, F.H., 2018. A comparative study of lstm and dnn for stock market forecasting. In: In 2018 IEEE International Conference on Big Data (Big Data), pp. 4148–4155.

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt. Sci. Rep. 11 (1), 1606.

Sparnaaij, L.D., 1959. The Analysis of Bunch Production in the Oil Palm. PhD thesis.

Tuerxun, A., Mohamed Shariff, A.R., Janius, R., Abbas, Z., Mahdiraji, G.A., 2020. Oil palm fresh fruit bunches maturity prediction by using optical spectrometer. IOP Conf. Series: Earth Environ. Sci. 540 (1), 012085.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177, 105709.

Verma, A., Dong, X., 2016. Detection of ventricular fibrillation using random forest classifier. J. Biomed. Sci. Eng. 09, 259–268.

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.

Whitmire, C.D., Vance, J.M., Rasheed, H.K., Missaoui, A., Rasheed, K.M., Maier, F.W., 2021. Using machine learning and feature selection for alfalfa yield prediction. AI 2 (1), 71–88.

Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., Zhu, M., Wu, X., 2019. Design of an integrated climatic assessment indicator (icai) for wheat production: A case study in Jiangsu province, China. Ecol. Indic. 101, 943–953.

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415, 295–316.