ORIGINAL RESEARCH

Predictions of oil volume in palm fruit and estimates of their ripeness: A comparative study of machine learning algorithms

Sherif Eneye Shuaib [1], Pakwan Riyapan [2],
Saysunee Jumrat [3,4], Yutthapong Pianroj [3,4],
Jirapond Muangprathub [3,4*]

[1]College of Digital Science, Prince of Songkla University, Songkla 90110, Thailand
[2]Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand
[3]Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand
[4]Integrated High-Value Oleochemical Research Center, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand

[*] To whom correspondence should be addressed. Email: jirapond.m@psu.ac.th

Abstract

Recent advances in precision agriculture highlight the crucial role of machine learning in predicting crop yields by discerning intricate patterns in agro-meteorological data. However, its adoption in the oil palm industry in Thailand remains limited. This study aimed to compare machine learning algorithms for predicting the oil content from different parts of both ripe and raw oil palm fruits (top, middle, and down). Additionally, we compared algorithms for predicting oil volume in semi-ripe and unripe fruits. Among the methods used, Random Forest and Gradient Boosting models mostly excelled in predicting the oil content at different positions on the oil palm fruit. In contrast, Decision Trees and XGBoost were the most accurate predictors of oil volume for semi-ripe and unripe oil palm fruits, respectively. Overall, this research emphasizes the potential of machine learning to enhance oil palm industry practices and optimize agricultural strategies in Thailand.

Keywords

fruit ripeness; machine learning; moisture content; oil palm industry; precision agriculture; supervised learning

1. Introduction

The *Elaeis guineensis Jacq.*, more widely known as the oil palm, is a single, unbranched tree that demands multiple years of dedication and hard work before it yields fresh fruit bunches with high oil content (Uning et al., 2020). Around one-third of global vegetable oil originates from this tree, surpassing the oil yield from soybean and rapeseed (Morcillo et al., 2013). Oil palm fruit offers two primary types of vegetable oil: crude palm oil from the fruit's mesocarp and palm kernel oil from its seeds. In optimal conditions, industrial oil palms generally yield between 12 and 18 tons of fruit annually per hectare. Notably, oil palms are unique permanent plants, deviating from the norm of annual, biennial, and perennial crops. Harvesting occurs bi-monthly throughout the plant's life, barring the early growth phase (Ismail & Mamat, 2002; Legros et al., 2009). Yields are relatively stable, albeit seasonal factors (Jelsma et al., 2019) and such issues as inadequate fertilization, pest infestation, and disease can affect them. Effective field management, disease control, and optimal harvesting patterns bolster yield quality and quantity. However, several factors can still reduce the overall yield (Gérard et al., 2017). Given these challenges, there is an increasing need for more precise cultivation techniques, where machine learning and data-driven

technologies can play a crucial role in optimizing field management and improving yield efficiency (Xia et al., 2024).

Thailand considers oil palm a pivotal economic crop due to its remarkable yield rate compared to other oil crops. In 2016, the nation dedicated 729,600 hectares to oil palm cultivation, which has been on an upward trajectory (Suppalakpanya et al., 2019). By 2020, Thailand's oil palm plantation area spanned over 9,954.27 km^2, trailing only Indonesia and Malaysia globally. The majority of this area, approximately 8,500 km^2, is situated in southern peninsular Thailand. Krabi boasts the largest plantation stretch in this region, whereas Ranong is renowned for its impressive yield per unit area (Worachairungreung et al., 2023). However, Thailand's oil palm industry faces challenges, and one of these challenges is the presence of over 200,000 small-scale growers. Additionally, associated production costs and logistical complexities further impact the industry. In contrast to key producers like Indonesia and Malaysia, which utilize vast transportation networks, Thai farmers predominantly employ smaller vehicles. Furthermore, sector-specific expertise is needed to boost the optimal amount of oil production. Thai oil palm cultivators face the pressing challenge of harvesting their crops prematurely, which diminishes oil content and reduces potential profits (Raksaseri, 2023; Treerutkuarkul, 2021). Machine learning can play an essential role in addressing the challenges of premature harvesting in Thailand's oil palm industry by predicting optimal harvesting times. Accurate prediction of when to harvest allows farmers to maximize oil content in the fruit, directly improving yields and increasing profitability. Given Thailand's reliance on small-scale growers and the logistical hurdles they face, adopting machine learning can streamline operations by providing data-driven insights on when to harvest, helping farmers avoid premature harvesting, which significantly lowers oil content.

With the advent of machine learning, precision agriculture is experiencing revolutionary advancements. These recent advancements have demonstrated the significance of machine learning in predicting crop yields by identifying linear and nonlinear patterns within intricate agro-meteorological data. Oil palm cultivation, striving to adhere to global sustainability benchmarks, also incorporates these innovations (Behmann et al., 2015; Chlingaryan et al., 2018; Dimitriadis & Goumopoulos, 2008; Rahman et al., 2018). Tasks like soil and crop management, crop selection, yield predictions, and more are evolving through machine learning. Despite this capability, the utilization of machine learning techniques for predictive analysis remains limited within the oil palm industry, particularly in Thailand. A systematic review by Khan et al. (2021) identified an imbalance in the focus of oil palm research, with 84% of studies using classification techniques and only a small proportion exploring regression methods for predictive analysis. The geographical distribution of the analyzed articles revealed that the top six countries involved in oil palm research were Malaysia, with 38 articles, followed by Indonesia, the UK, the USA, Australia, China, and Thailand, with 11, 8, 5, 4, and 4 articles, respectively. Based on the geographical distribution of studies, Thailand is significantly underrepresented compared to key players like Malaysia, further highlighting a gap in research specific to the Thai oil palm industry. The limited adoption of regression-based approaches for oil palm predictions has left a crucial gap in developing precision agriculture solutions for this crop. Hence, adopting automation and precision approaches through machine learning in this domain is imperative, especially in Thailand, as the integration of machine learning techniques can assist in guiding farmers on the best harvesting times to optimize oil yields. Such guidance can augment oil quality, aligning it with countries like Malaysia and Indonesia, which are considered premium producers.

This study aimed to compare machine learning algorithms for predicting the oil content from different parts of both ripe and raw oil palm fruits (top, middle, and down). Additionally, we compared algorithms for predicting oil volume in semi-ripe and unripe fruits. The findings have the potential to provide critical insights to farmers, enabling more informed decisions about harvesting times, thus improving oil quality and stabilizing supply-demand relationships. This research not only contributes to the growing body of work in precision agriculture but also underscores the need for Thailand to catch up with global leaders in oil palm cultivation by leveraging the full potential of machine learning.

2. Materials and methods

Machine learning is becoming increasingly pervasive, impacting human society and the natural world, including agriculture (Dahal et al., 2021; Gonzalez-Rivero et al., 2020). These algorithms enable data-based decision-making and predictions. Predicting agricultural outcomes, such as crop yields, can significantly affect food security and livelihoods.

Machine learning algorithms are typically categorized into three main types based on their learning feedback (Murphy, 2018; Verbraeken et al., 2020): supervised, unsupervised, and reinforcement learning. Our study focused on supervised learning, which predicts outcomes based on labeled data. Moreover, this approach has several algorithms that can be used to build the predictive model. Thus, to predict agricultural outcomes, this study selected seven popular algorithms based on different constructions to fit the model, namely *Linear Regression*, *Decision Trees*, *Support Vector Regression*, *Random Forest*, *Gradient Boosting*, *Extra Gradient Boosting Machine*, and *Light Gradient Boosting Machine*.

2.1. Linear regression

Simple linear regression involves predicting a dependent variable using a single independent variable, known as univariate regression analysis. In simple linear regression, the dependent and independent variables are differentiated to determine the relationship between the two variables. This relationship is similar to correlation, but unlike correlation, simple linear regression distinguishes between the dependent and independent variables.

2.2. Decision Trees

Decision Trees (DT) are a non-parametric and simple structure classification algorithm for capturing nonlinear relationships between features and classes (Friedl & Brodley, 1997). They are represented as a tree-based hierarchy of rules and functions by recursively partitioning or splitting the input data into smaller subsets (Friedl & Brodley, 1997; Song & Ying, 2015). This splitting process is guided by thresholds defined at each internal node in the tree. Starting from the root node in the DT, the input data are successively divided into sub-nodes and further sub-nodes (Sharma et al., 2013; Song & Ying, 2015). Ultimately, the input data are classified based on this binary subdivision, with the final nodes, called leaf nodes or leaves, representing the target classes (Maxwell et al., 2018; Pal & Mather, 2003). Despite their effectiveness, DTs have limitations. They may not always produce optimal solutions, as they rely on a single tree. Overfitting is also a common issue with DTs, requiring careful consideration during their use.

2.3. Support Vector Regression

Support Vector Regression (SVR) is a supervised learning model for classification and regression. This approach is beneficial for examining the connections between a dependent variable and one or more independent variables. By framing an optimization problem, SVR learns a regression function that connects input predictor variables to the output observed response values. SVR is advantageous as it balances model complexity and prediction error, demonstrating strong performance, particularly with high-dimensional data.

Let the dot product pace $\mathfrak{R}^d$ be our data universe with vectors $x \in \mathfrak{R}^d$ as objects and S be a sample set such that $S \subset \mathfrak{R}^d$. Suppose $f : \mathfrak{R}^d \to \mathfrak{R}$ is the target function and $D = \{(x, y) \mid x \in S \text{ and } y = f(x)\}$ is the training set. The regression problem is to find the best-approximated model $\tilde{f} : \mathfrak{R}^d \to y$ for the true underlying function f mapping input "x" to output "y" by using D such that $\tilde{f}(x) \cong f(x)$. SVR was mainly developed to solve the nonlinear regression problem, which is more challenging than linear regression problems.

2.4. Random Forest

Random Forest (RF) is a robust ensemble learning method that utilizes multiple decision tree classifiers to overcome the limitations of a single classifier in achieving the optimal solution (Belgiu & Drăguţ, 2016; Breiman, 2001; Cutler et al., 2007). By incorporating many trees instead of a single tree, the RF algorithm employs a majority vote technique to assign a final class label, thereby improving accuracy. This approach also helps address issues related to handling many variables in the model. To achieve this, each tree in the RF model is trained on a randomly generated subset of the training data and uses only a subset of the tree's variables. While this strategy may reduce the performance of individual trees, it reduces the correlation between trees, making the ensemble more reliable. Additionally, since RF incorporates multiple classifiers, there is no need to prune individual trees, simplifying the model-building process (Breiman, 2001). Overall, RF is effective in improving classification performance and handling complex datasets.

2.5. Gradient Boosting Machines

Gradient Boosting Machine (GBM) is a prediction algorithm based on decision trees (Friedman, 2001). It constructs a model gradually, additively, and sequentially by combining multiple decision trees in linear combinations (Biau et al., 2019). GBM shares a foundational concept with AdaBoost, making it easier to understand. However, the critical difference lies in how they address the weaknesses of weak classifiers.

In AdaBoost, weaknesses are identified using high-weight data points that are difficult to fit, whereas in GBM, they are identified using gradients. The methodology involves modeling data with simple base classifiers, analyzing errors, focusing on hard-to-fit data points to correct them, and assigning weights to each predictor to combine all predictions for a final result.

GBM has demonstrated significant success in various applications, including text classification, web searching, landslide susceptibility assessment, and image classification (Chen & Guestrin, 2016; Samat et al., 2020). However, GBM may not perform well with exceptionally noisy data, as it can lead to overfitting (Jafarzadeh et al., 2021).

2.6. XGBoost

Extreme Gradient Boosting Machine (XGBoost), developed by Chen and Guestrin (2016), is a gradient tree boosting method that builds upon regular Gradient Boosting Machines (GBM) with several enhancements. It introduces features like regularization to prevent overfitting tree pruning that specifies tree depth using the Maximum Tree Depth (MTD) parameter and prunes the tree backward instead of based on loss criteria, resulting in improved computational performance and parallelism that utilizes a block structure for parallel learning, leading to faster computation (Zhong et al., 2022). XGBoost employs a decision tree as a booster and has shown outstanding performance across various ranking tasks, classification, and regression (Samat et al., 2020). Despite its success, XGBoost has not been extensively studied in remote sensing image classification tasks, particularly concerning spectral and spatial features, classification accuracy, computational efficiency, and the influence of crucial parameters. The main advantages and disadvantages of only numerical values are accepted for processing.

2.7. LightGBM

The LightGBM algorithm is a gradient-boosting framework that builds on the concept of decision trees (Shi et al., 2019). It is designed to reduce computation time while maintaining high accuracy (Friedman, 2001; Ke et al., 2017; Shi et al., 2019). An important distinction between LightGBM and other decision-tree-based algorithms lies in its tree growth strategy. While traditional methods grow trees level-wise (horizontally), LightGBM grows trees leaf-wise (vertically), leading to a more complex structure (Machado et al., 2019). This approach enhances the efficiency of the algorithm. Although the implementation of XGBoost and LightGBM is similar, Light-

GBM outperforms XGBoost in training speed and ability to handle large datasets. However, both methods require extensive parameter tuning for optimal classification performance and reliable results compared to methods like Random Forest (RF).

2.8. Data preparation

The dataset used in this study was obtained from the Southern Palm Oil Industry (1993) Co. Ltd. in Surat Thani province, Thailand. The data collection process involved two main components. Firstly, data were gathered in real-time situations using specially designed devices based on IoT (Internet of Things) technology. These devices featured Resistive Level Sensors and Temperature Humidity Sensors installed on the ESP32 NodeMCU board, as depicted in Figure 1.
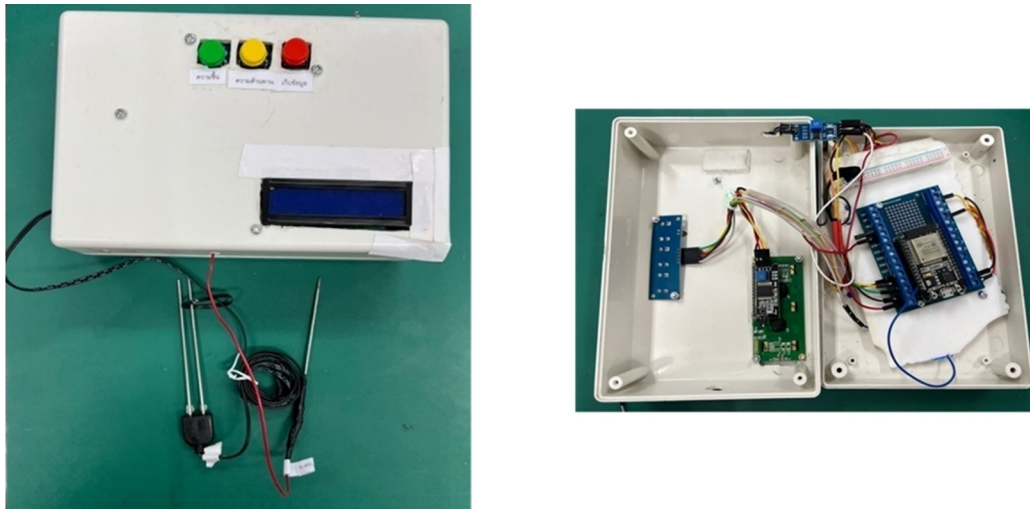
Figure 1 IoT devices for collecting oil palm data.

The collected humidity and resistance level data were transmitted to the server using the Internet as a web application. Simultaneously, laboratory experiments were conducted utilizing Soxhlet extraction (see Figure 2), followed by laboratory scientists performing the measurements and recording the results in a spreadsheet. Due to the preliminary nature of this study and the time constraints involved, a sample size of 30 seedlings was used. The data collection spanned two months, from February to March 2023, with data collected continuously daily throughout this period. Each day, the testing team could test only 2–3 samples. Additionally, to ensure data accuracy and completeness, efforts were made to avoid missing values, especially considering the limited sample collection.

For the IoT data, the average of measurements from 3 positions at each level of the oil palm fruit was used to minimize potential discrepancies. The data source in Surat Thani province of Thailand provided valuable insights for the study's analysis and conclusions.

2.9. Data model

Machine learning involves a set of methods where computers can model the relationship between numerical data representations and specific target values (Hao & Ho, 2019). Machine learning is broadly categorized into two types: supervised learning (which trains on known inputs and outputs to predict future outputs) and unsupervised learning (which discovers patterns or structures in the input data) (Swamynathan, 2017) (Figure 1). The machine learning techniques applied in our study included linear (*linear Regression*) and nonlinear models (*Decision Tree, Random Forest, Gradient Boosting (GB), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Support Vector Regression (SVR)*).

Figure 2 Soxhlet extraction.

These algorithms are some of the most commonly utilized approaches in recent literature (Çakıt & Dağdeviren, 2022). Each belongs to a distinct algorithmic family with fundamentally different internal architectures (Fernández-Delgado et al., 2014). Basic descriptions of these methods are presented in Table 1.

All the models were implemented in Python/Jupyter Notebook, and scikit-learn packages were employed to incorporate these machine-learning algorithms into our study (Pedregosa et al., 2011). An overview of the modeling done is presented in Figure 3.

2.10. Performance metrics

To assess the disparity between the observed and predicted values (error) for the models employed in this study, various performance metrics were utilized, as described by Çakıt and Karwowski (2017) and Çakıt et al. (2020). In this analysis, the effectiveness of the algorithms used was determined by evaluating the model's accuracy through metrics, such as Root Mean Squared Error (*RMSE*), Mean Squared Error (*MSE*), and the coefficient of determination (R^2). Low values of *RMSE* and *MSE* indicate more precise model outcomes, while higher R^2 values signify a stronger alignment between observed and estimated values. These metrics were computed using the following equations:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(P_i^v - A_i^v\right)^2 \tag{1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i| \tag{2}$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} \left(P_i^v - A_i^v\right)^2}{\sum_{i=1}^{n} A_i^{v^2}}\right) \tag{3}$$

Table 1 Overview of machine learning methods and Python packages.

Method	Description	Python package
Linear Regression	Simple statistical method modeling the relationship between dependent and independent variables	scikit-learn
Decision Tree	Tree-like model making decisions based on input features, suitable for classification/regression	scikit-learn
Random Forest	Ensemble method building multiple decision trees to improve accuracy and reduce overfitting	scikit-learn
Gradient Boosting	Ensemble method building trees sequentially, correcting errors of previous ones	scikit-learn
XGBoost	Efficient and scalable implementation of gradient boosting, known for speed and performance	xgboost
LightGBM	Gradient boosting framework designed for speed and efficiency, suitable for large datasets	lightgbm
Support Vector Regression (SVR)	Regression technique using support vector machines to find a hyperplane representing the relationship	scikit-learn
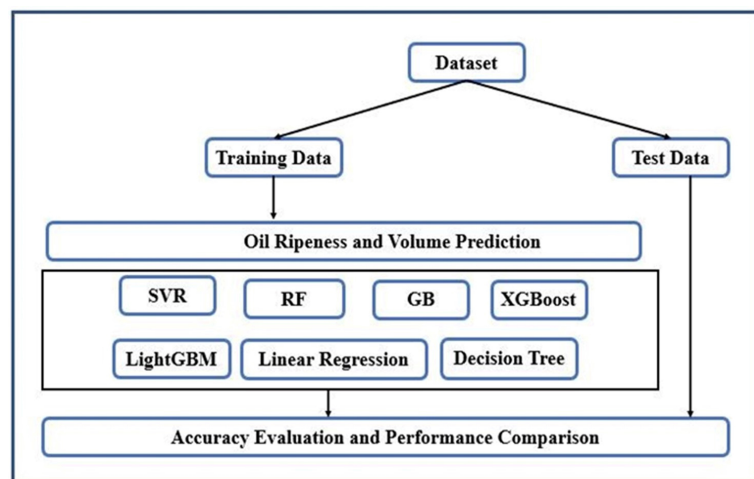
Figure 3 Overview of the supervised machine learning models used in this study.

where

- A_i^v and P_i^v are the actual and predicted values, respectively,
- e_i: is the prediction error for each seedling.
- n: total number of seedlings tested.
- $i = 1, 2, 3, \ldots, n$.

3. Results

In this study, we present the results obtained from our predictive analysis. The predictions of oil content in ripe and raw oil palm fruit at the top, middle, and down positions are illustrated in Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, and Figure 9, respectively.

Additionally, Figure 10 and Figure 11 present the results obtained from oil volume predictions in semi-ripe and unripe oil palm fruit, respectively.

4. Discussion

4.1. Performance evaluation based on oil palm fruit position (ripe fruit)

In our analysis of oil content prediction in ripe fruit (Table 2), in the 'Top' position, both GBM and RF models exhibited superior predictive accuracy compared to the
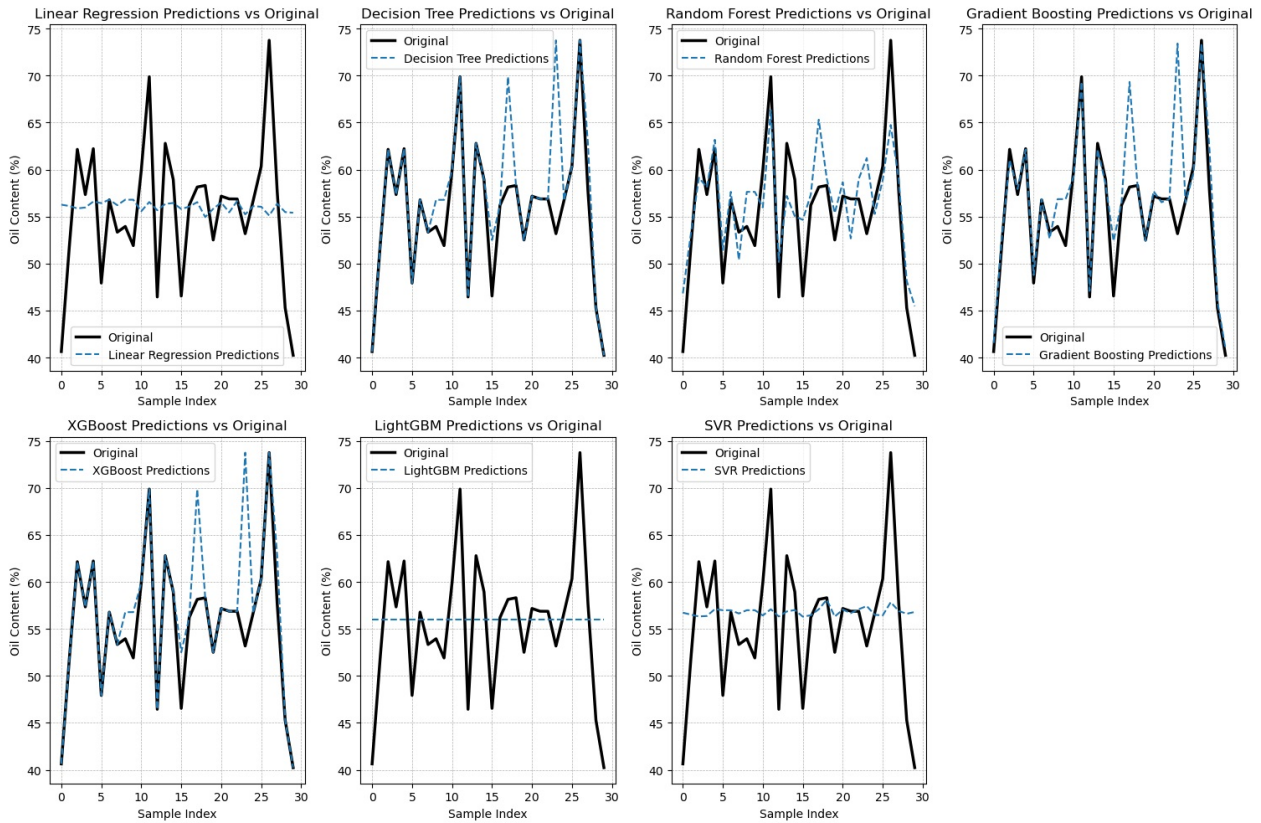
Figure 4 Comparison between real-world observations and the outputs generated by the proposed algorithms for ripe oil palm in the top position.
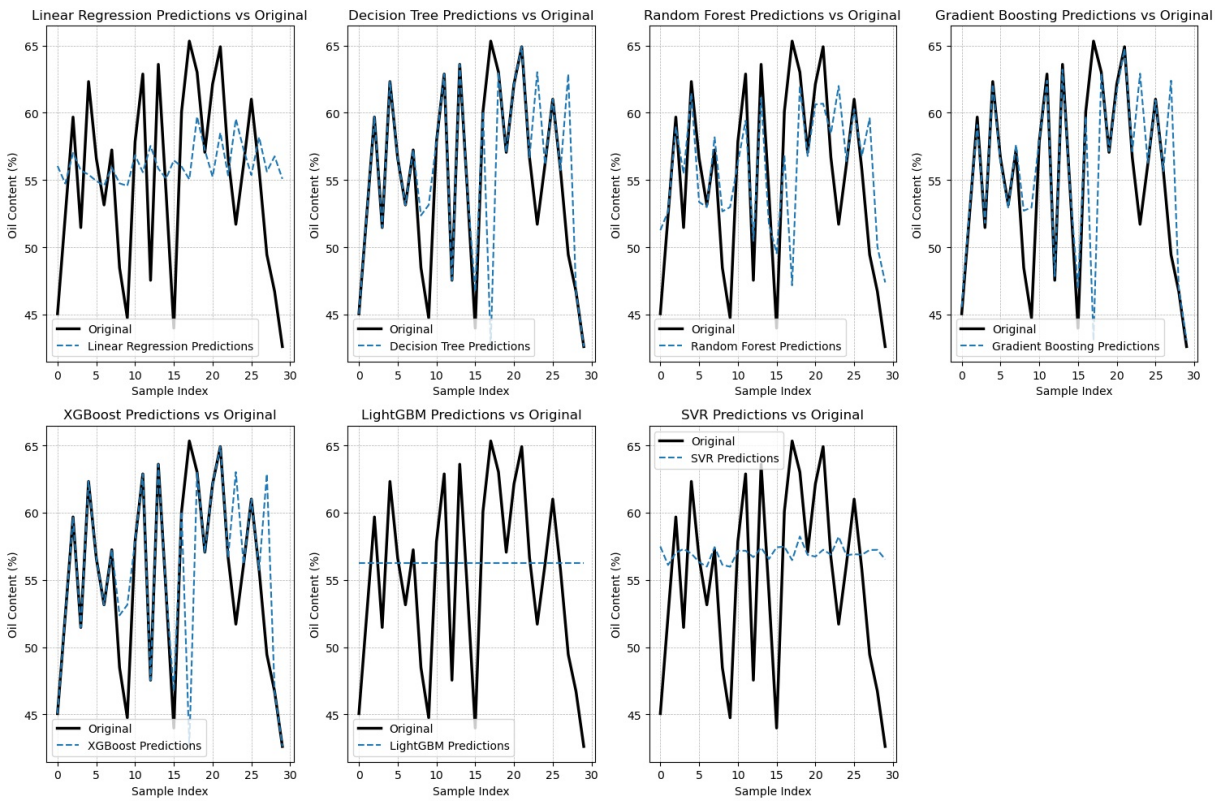
Figure 5 Comparison between real-world observations and the outputs generated by the proposed algorithms for ripe oil palm in the middle position.

Figure 6 Comparison between real-world observations and the outputs generated by the proposed algorithms for ripe oil palm in the down position.
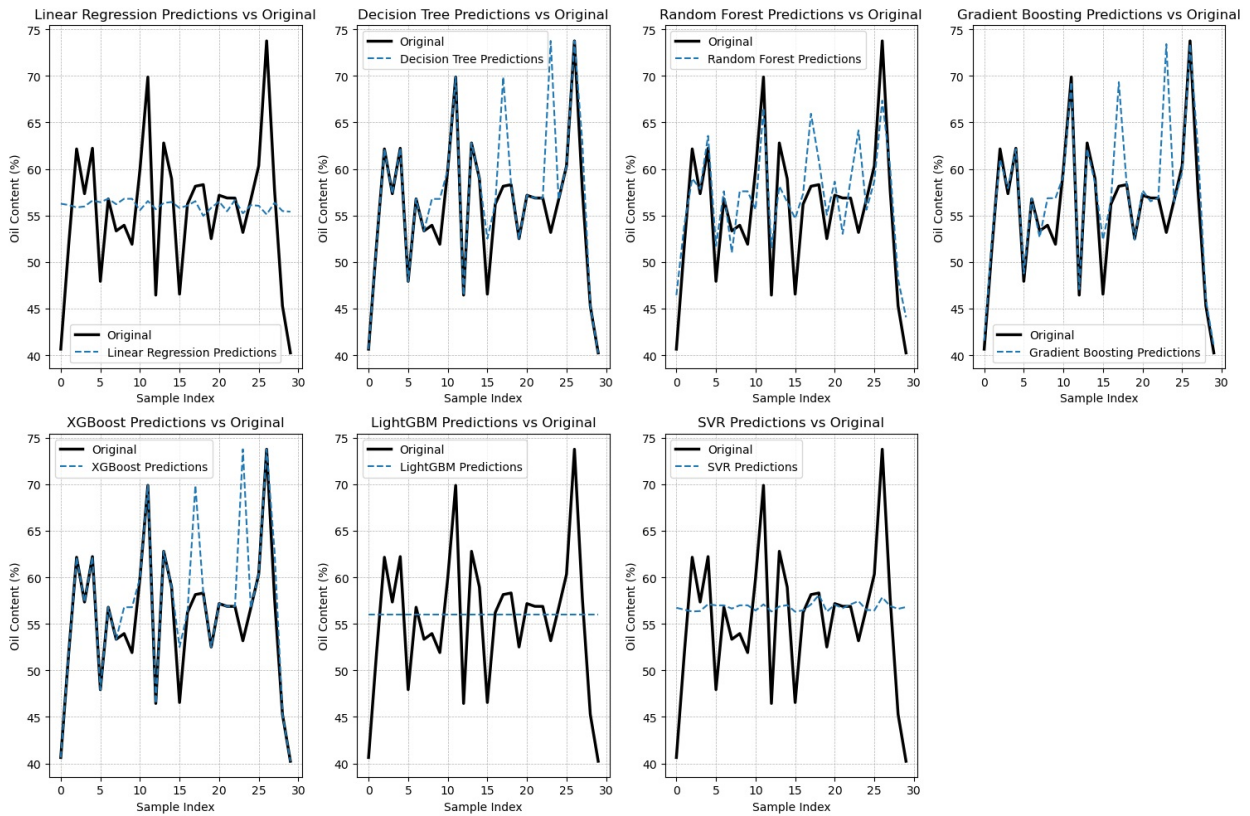
Figure 7 Comparison between real-world observations and the outputs generated by the proposed algorithm for raw oil palm in the top position.
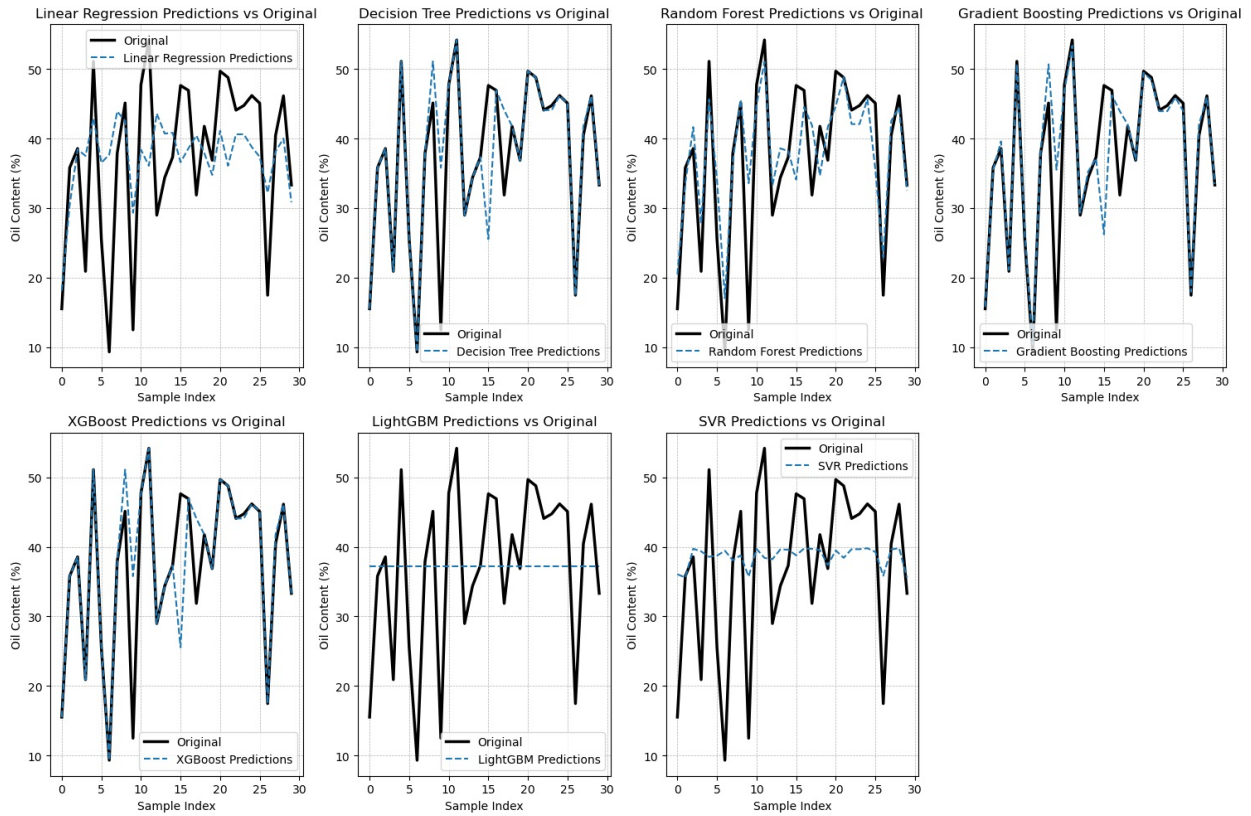
Figure 8 Comparison between real-world observations and the outputs generated by the proposed algorithms for raw oil palm in the middle position.
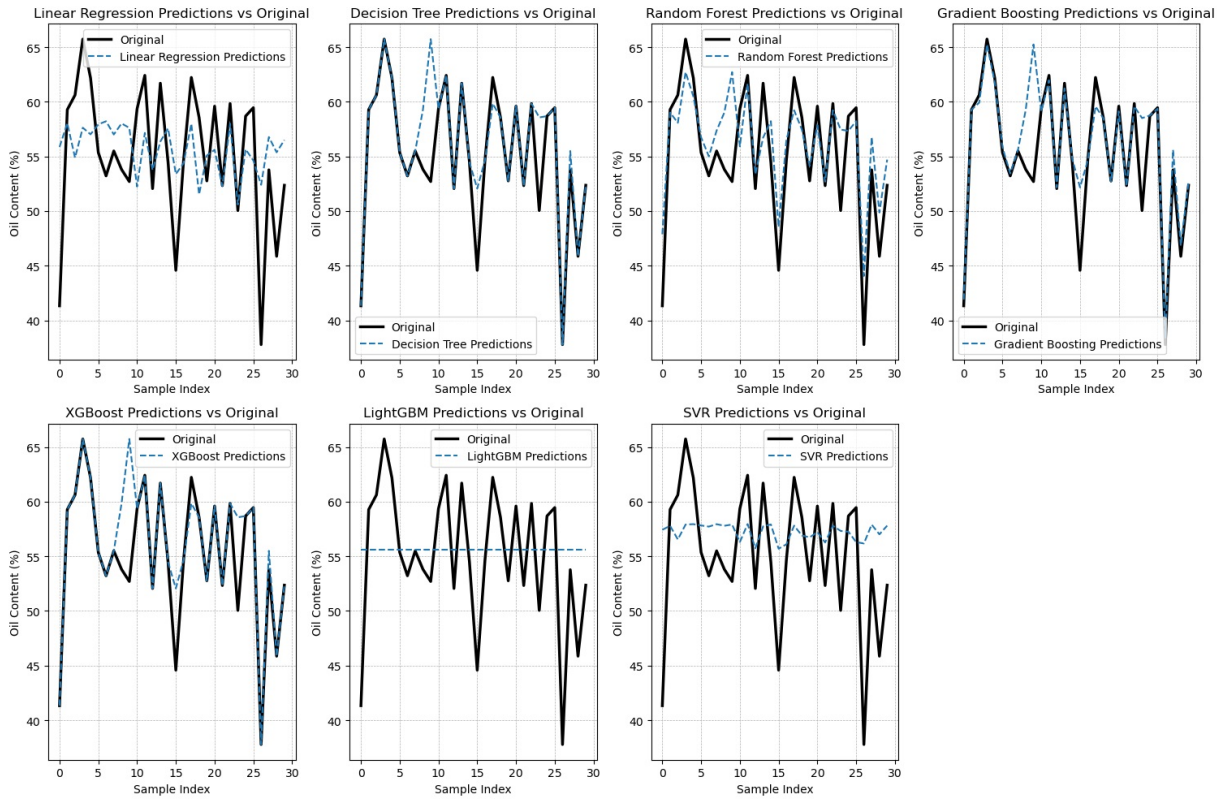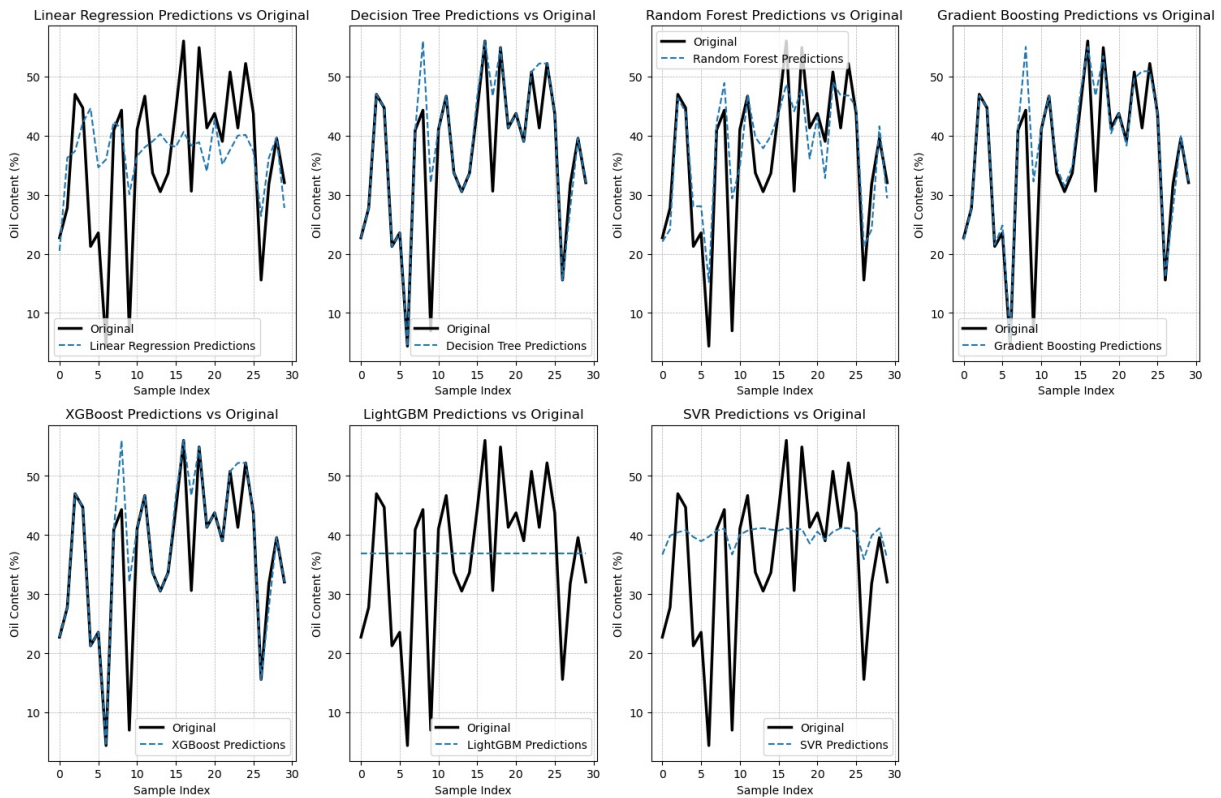

Figure 9 Comparison between real-world observations and the outputs generated by the proposed algorithms for ripe oil palm in the down position.

Figure 10 Comparison between real-world observations and the outputs generated by the proposed algorithm for semi-ripe oil palm.

Figure 11 Comparison between real-world observations and the outputs generated by the proposed algorithm for unripe oil palm.

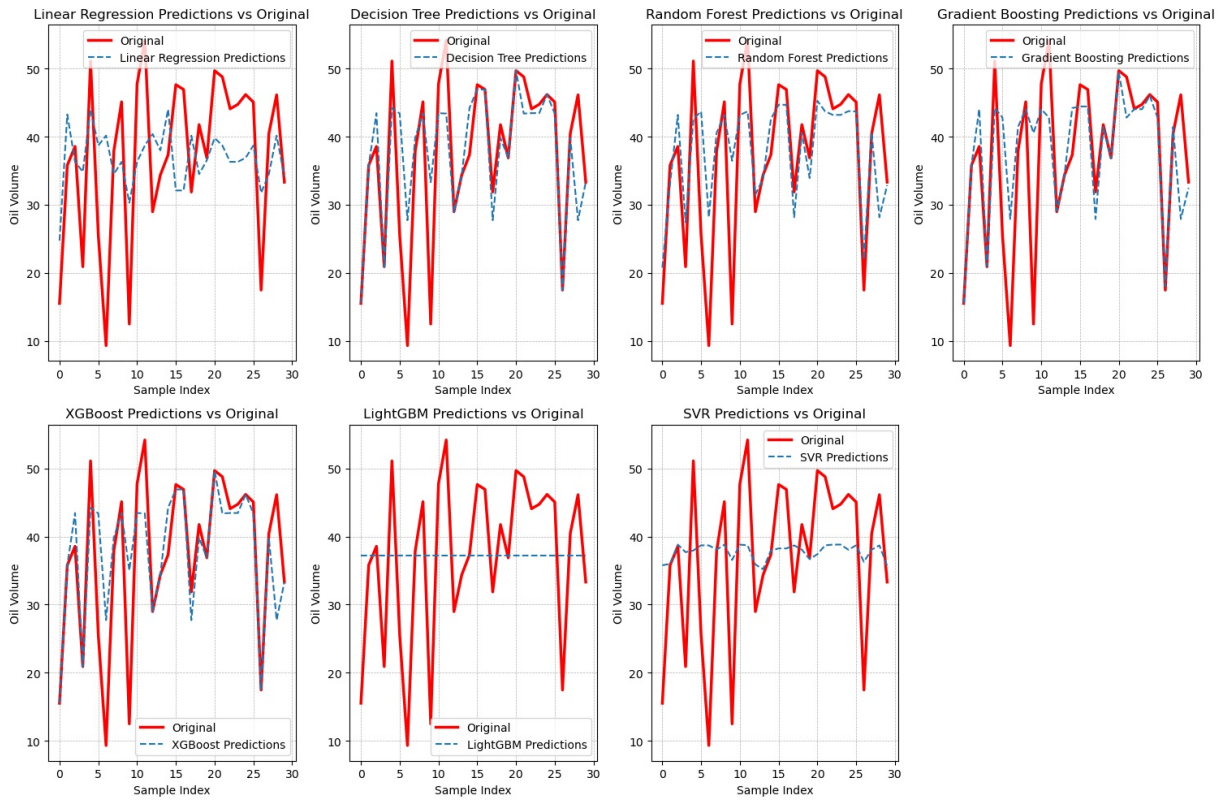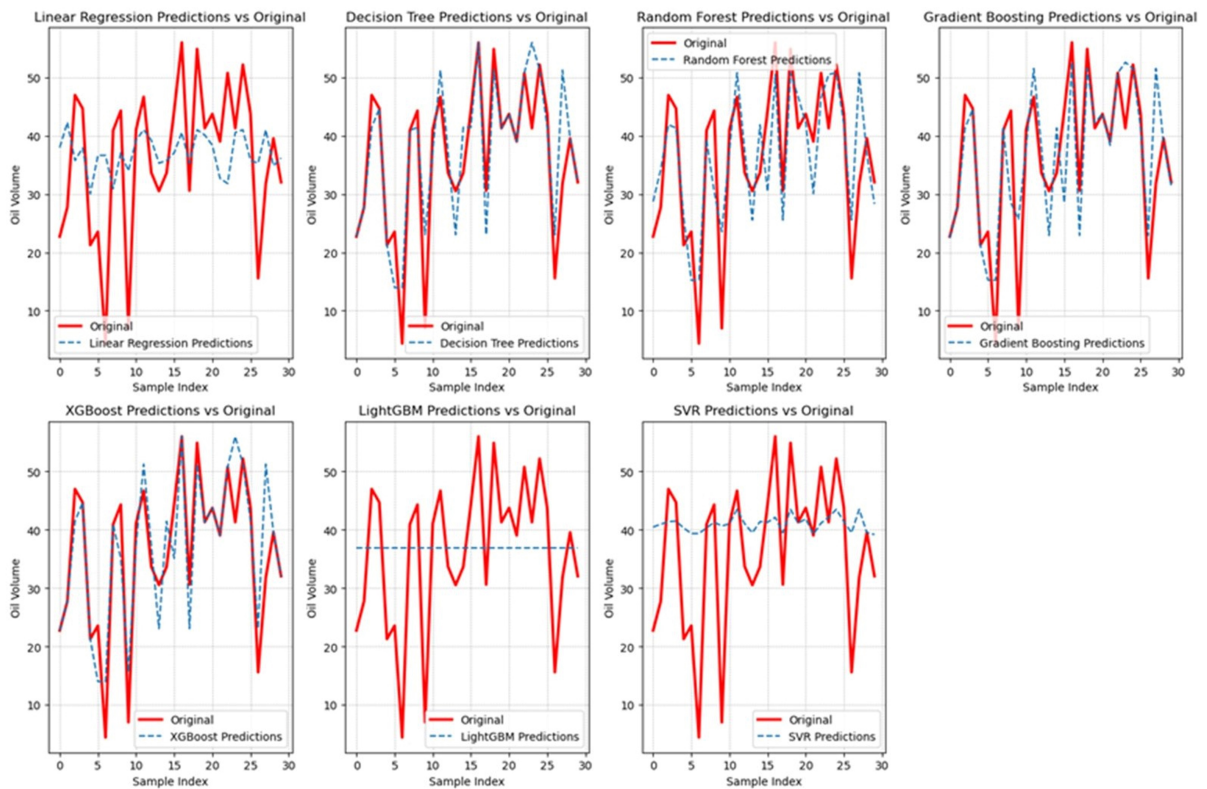**Table 2** Performance metrics for oil content prediction.

s/no	Model	MSE	MAE	R^2
		Ripe Fruit (Top)		
1	Linear Regression	53.2207	5.3482	−0.0008
2	Decision Tree	21.7890	1.7010	0.5903
3	Random Forest	19.8729	3.6296	0.6263
4	Gradient Boosting	20.9668	2.0506	0.6057
5	XGBoost	21.7806	1.7035	0.5904
6	LightGBM	53.4167	5.4062	−0.0045
7	SVR	52.7411	5.1968	0.0082
		Ripe Fruit (Middle)		
1	Linear Regression	44.3771	5.5154	−0.0021
2	Decision Tree	30.5705	2.0820	0.3097
3	Random Forest	28.0404	3.6351	0.3668
4	Gradient Boosting	29.4955	2.2857	0.3339
5	XGBoost	30.5648	2.0830	0.3098
6	LightGBM	45.5526	5.5717	−0.0287
7	SVR	46.4804	5.4742	−0.0496
		Ripe Fruit (Down)		
1	Linear Regression	35.5574	4.7928	0.1246
2	Decision Tree	11.2403	1.2877	0.7233
3	Random Forest	13.5343	2.9764	0.6668
4	Gradient Boosting	11.0222	1.5340	0.7286
5	XGBoost	11.4525	1.3073	0.7180
6	LightGBM	40.9264	4.9871	−0.0076
7	SVR	42.2113	5.0968	−0.0392

other algorithms. The RF model, in particular, had a higher R^2, indicating a more precise fit to the data. The XGBoost and DT methods closely followed suit, underscoring their strong performance in capturing variations in ripeness. Similarly, the Decision Tree and RF models had the lowest MSE and MAE values for oil palm fruit in the 'Middle' position. GBM and XGBoost also demonstrated strong performance, emphasizing their reliable predictive capabilities across different positions on the palm tree. Lastly, in the 'Down' position, the Decision Tree and Gradient Boosting models consistently outperformed the other methods. These findings align with Sinambela et al. (2020), who also highlighted the central position of the fruit as a critical indicator of ripeness. Sinambela's statistical approach using two-way ANOVA supports our conclusion that model accuracy is influenced by the fruit position on the tree.

4.2. Performance evaluation based on oil palm fruit position (raw fruit)

For oil content prediction in raw fruit (Table 3), the RF model stood out as the top-performing model in the 'Top' position, demonstrating superior accuracy with the lowest MSE values and a relatively high R-squared value, indicating a strong fit to the data. The other models, such as GBM, XGBoost, and DT, performed well with slightly higher MSE and a lower R-squared value than the RF model.

In the 'Middle' position, the DT model with the lowest MAE value and GBM with the lowest MSE and the highest R-squared value demonstrated superior performance compared to the other models. This indicates their ability to predict oil content in raw fruit more accurately than the other models. The Random Forest and XGBoost models also showed competitive performance in this position, highlighting their consistency across different scenarios.

In the 'Down' position, DT, with the lowest MAE value, and GBM, with both the lowest MSE and the highest R-squared value, outperformed the other models.

Table 3 Performance metrics for oil content prediction.

s/no	Model	MSE	MAE	R^2
	Raw Oil Palm (Top)			
1	Linear Regression	53.2207	5.3482	−0.0008
2	Decision Tree	21.7890	1.7010	0.5903
3	Random Forest	19.8787	3.6780	0.6262
4	Gradient Boosting	20.9668	2.0506	0.6057
5	XGBoost	21.7806	1.7035	0.5904
6	LightGBM	53.4167	5.4062	−0.0045
7	SVR	52.7411	5.1968	0.0082
	Raw Oil Palm (Middle)			
1	Linear Regression	107.9818	8.4360	0.2432
2	Decision Tree	40.6303	2.1863	0.7153
3	Random Forest	44.8086	4.8512	0.6860
4	Gradient Boosting	39.2065	2.5192	0.7252
5	XGBoost	40.6236	2.1873	0.7153
6	LightGBM	142.6922	9.5858	−0.0045
7	SVR	105.2121	8.1649	0.0829
	Raw Oil Palm (Down)			
1	Linear Regression	127.0705	8.6701	0.2244
2	Decision Tree	38.8223	2.3367	0.7630
3	Random Forest	45.5283	5.0568	0.7221
4	Gradient Boosting	38.0251	2.7397	0.7672
5	XGBoost	38.8184	2.3376	0.7631
6	LightGBM	164.3930	10.3114	−0.0034
7	SVR	155.3703	9.2831	0.0517

The XGBoost model also performed well, indicating its robust predictive capabilities even in challenging scenarios. These results highlight the effectiveness of DT and GBM for predicting oil content in raw fruit, with XGBoost also proving to be a strong performer. Similar approaches were utilized by Sae-Tang (2020), who employed machine learning models, like convolutional neural networks, to estimate oil content in fresh fruit bunches using surface color as a predictor. Their results also demonstrated high accuracy, emphasizing the importance of advanced algorithms in oil content prediction.

4.3. Performance evaluation based on oil volume (semi-ripe fruit)

In our evaluation of predictive models for oil volume in semi-ripe fruit, Table 4 compares the seven algorithms used, highlighting their accuracy and performance. Linear Regression demonstrated the weakest predictive capability, with the highest MSE of 126.5536 and an MAE of 9.5926, signaling significant deviations from actual values. The low R-squared value of 0.1130 underscores its poor fit to the data, making it an ineffective model for the current data. In contrast, DT emerged as the best-performing model, with the lowest MSE (58.2306) and MAE (4.3087) and an R-squared value of 0.5919, suggesting it can effectively capture patterns in the data for predicting oil volume in semi-ripe oil palm fruit. While Random Forest (RF) also performed well, it did not surpass DT, showing a higher MSE of 69.1243 and a lower R-squared value of 0.5156. This indicates that although RF remains a strong model, it does not provide the same level of accuracy as DT. SVR yielded high prediction errors, with an MSE of 131.8618 and a weak R-squared value of 0.0759, marking it as less reliable than the other models. GBM showed moderate performance, with an MSE of 70.6910 and a fair fit (R-squared = 0.5046), while XGBoost performed better, with a lower MSE (60.7034) and an R-squared value of 0.5746, making it a competitive alternative to DT. Finally, LightGBM displayed poor predictive capability, with the

Table 4 Performance metrics for predictive models on semi-ripe and unripe fruit.

Type	Method	MSE	MAE	R^2
		Semi Ripe		
1	Linear Regression	126.5536	9.5926	0.1130
2	Decision Trees	58.2306	4.3087	0.5919
3	Random Forest	69.1243	5.6283	0.5156
4	SVR	131.8618	8.8121	0.0759
5	Gradient Boosting	70.6910	4.7757	0.5046
6	XGBoost	60.7034	4.7757	0.5746
7	LightGBM	142.6922	9.5858	−0.0000
		Unripe		
1	Linear Regression	150.6958	9.8980	0.0802
2	Decision Trees	46.0533	4.3287	0.7189
3	Random Forest	66.1267	6.6388	0.5964
4	SVR	168.8536	9.2666	−0.0306
5	Gradient Boosting	59.8810	5.2948	0.6345
6	XGBoost	45.0934	4.5053	0.7248
7	LightGBM	164.3930	10.3114	−0.0034

highest MSE (142.6922) and an R-squared value close to zero, demonstrating its unsuitability for this analysis. In conclusion, DT stands out as the most accurate model for predicting oil content in semi-ripe fruit, outperforming all the other algorithms. This complements the findings by Wangrakdiskul and Yodpijit (2015), who applied an exponential growth model for forecasting palm oil production and consumption in Thailand, showing the significance of accurate prediction models in optimizing palm oil yield.

4.4. Performance evaluation based on oil volume (unripe fruit)

In our evaluation of predictive models for oil volume in unripe fruit, Table 4 highlights distinct variations in accuracy and performance across the seven algorithms. Linear Regression demonstrated the weakest performance, with an MSE of 150.6958, reflecting a considerable discrepancy between the predicted and actual values. Its MAE of 9.8980 and low R-squared value of 0.0802 further emphasize its limited capacity to capture the underlying patterns in the data. In contrast, DT significantly outperformed Linear Regression, with a notably lower MSE of 46.0533 and MAE of 4.3287, indicating a more accurate prediction of oil volume in unripe fruit. The R-squared value of 0.7189 underscores the model's ability to fit the data well. RF exhibited a slightly higher MSE of 66.1267 compared to DT. However, it still provided a solid fit with an R-squared value of 0.5964, positioning it as a robust alternative for prediction. SVR performed poorly, yielding an MSE of 168.8536 and a negative R-squared value (−0.0306), signifying its lack of predictive accuracy and weak model fit. GBM displayed moderate predictive capability, with an MSE of 59.8810 and an R-squared value of 0.6345. XGBoost further enhanced performance, delivering a lower MSE of 45.0934 and the highest R-squared value (0.7248) among the models evaluated. LightGBM, however, showcased poor performance, with an MSE of 164.3930 and an R-squared value of −0.0034, reflecting its inability to model the data effectively. Overall, DT and XGBoost emerged as the top performers in predicting oil volume in unripe fruit. These results resonate with the work of Puttinaovarat and Horkaew (2019), who leveraged deep learning and machine learning methods for identifying oil palm plantations, demonstrating the versatility and robustness of advanced machine learning techniques in the context of oil palm yield prediction and detection.

Based on the discussions of the obtained results, the following are suggestions for the models employed in this study to improve the current understanding of oil palm fruit harvesting and processing. Firstly, DT, XGBoost, and RF consistently demonstrate a strong predictive accuracy for oil content across different stages of ripeness.

The accurate predictions of oil content help determine the optimal time for harvesting based on the fruit ripeness. This will reduce waste generated from harvesting fruit too early or too late, directly improving yields. Thus, industry professionals can implement data-driven harvesting strategies, ensuring that only fruit at peak ripeness is harvested to reduce losses, streamline operations, and increase profitability significantly.

Next, the analysis of fruit positions (Top, Middle, Down) also reveals that certain models perform better for specific positions on the palm tree. This implies that harvesting strategies can be refined based on the position of the fruit bunches, leading to more precise and customized harvesting methods. Professionals can tailor harvesting schedules to focus on specific areas of the tree where fruit shows the highest potential for oil yield, enhancing labor efficiency and reducing the environmental impact of broad harvesting approaches.

Lastly, machine learning models, especially when embedded into automated systems, can provide real-time predictions for oil content in different oil palm fruit positions. This can be paired with IoT devices or remote sensors, enabling continuous monitoring and data collection for model inputs. Large-scale plantations can leverage these models to monitor oil content continuously, allowing for dynamic adjustment of harvest schedules and predictive maintenance of processing facilities. This approach is consistent with the findings of Worachairungreung et al. (2023), who employed machine learning and data fusion techniques for classifying land use and land cover in oil palm plantations. Furthermore, Suppalakpanya et al. (2019) applied exponential time-series methods for forecasting crude palm oil prices and production in Thailand, which can complement the models we have discussed, opening avenues for more integrated forecasting approaches.

5. Conclusions

Recent advancements in precision agriculture have highlighted the critical role of machine learning in predicting crop yields, leveraging its ability to discern intricate linear and nonlinear patterns within agrometeorological data. Despite this potential, the adoption of machine learning techniques for predictive analysis remains limited, especially within the oil palm industry in Thailand. While prior studies have explored various aspects of oil palm ripeness through diverse methods, this research builds upon that foundation by focusing on predicting oil ripeness based on moisture content in different oil palm positions (top, middle, and down) and oil volume content for semi-ripe and raw data, utilizing a supervised learning approach.

In the top position, both XGBoost and RF models demonstrated superior accuracy compared to the other methods. The Decision Tree and Random Forest models exhibited the lowest MSE and MAE values in the middle position. In contrast, in the down position, the Decision Tree and Gradient Boosting models consistently outperformed the other methods. DT and GBM demonstrated exceptional raw oil volume prediction accuracy with low MSE, MAE, and a high R-squared value. In contrast, XGBoost emerged as the standout model for oil volume content, proving to be the most accurate predictor for semi-ripe and unripe oil palms compared to the other methods.

This study highlights the potential of machine learning to enhance oil palm industry practices. It reveals top-performing models for accurate ripeness and volume predictions and provides valuable insights for optimizing agricultural strategies in Thailand.

Data Availability

The data that support the findings of this study is available at: https://figshare.com/articles/dataset/Predictions_of_oil_volume_in_palm_fruit_and_estimates_of_their_ripeness_A_comparative_study_of_machine_learning_algorithms/28090769

Code Availability

The code used to analyze the data in the current study is available at: https://github.com/Ses4short/Oil-palm-Thailand

Acknowledgments

References

Behmann, J., Mahlein, A. K., Rumpf, T., Römer, C., & Plümer, L. (2015). A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, *16*, 239–260. https://doi.org/10.1007/s11119-014-9372-7

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine Learning*, *108*, 971–992. https://doi.org/10.1007/s10994-019-05787-1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Çakıt, E., & Dağdeviren, M. (2022). Predicting the percentage of student placement: A comparative study of machine learning algorithms. *Education and Information Technologies*, *27*(1), 997–1022. https://doi.org/10.1007/s10639-021-10655-4

Çakıt, E., & Karwowski, W. (2017). Predicting the occurrence of adverse events using an adaptive neuro-fuzzy inference system (ANFIS) approach with the help of ANFIS input selection. *Artificial Intelligence Review*, *48*(2), 139–155. https://doi.org/10.1007/s10462-016-9497-3

Çakıt, E., Karwowski, W., & Servi, L. (2020). Application of soft computing techniques for estimating emotional states expressed in Twitter® time series data. *Neural Computing and Applications*, *32*(8), 3535–3548. https://doi.org/10.1007/s00521-019-04048-5

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Dahal, S., Schaeffer, R., & Abdelfattah, E. (2021). Performance of different classification models on national coral reef monitoring dataset. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0662–0666). IEEE. https://doi.org/10.1109/CCWC51732.2021.9376135

Dimitriadis, S., & Goumopoulos, C. (2008). Applying machine learning to extract new knowledge in precision agriculture applications. In *2008 Panhellenic Conference on Informatics* (pp. 100–104). IEEE. https://doi.org/10.1109/PCI.2008.30

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(33), 399–409. https://doi.org/10.1016/S0034-4257(97)00049-7

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Gérard, A., Wollni, M., Hölscher, D., Irawan, B., Sundawati, L., Teuscher, M., & Kreft, H. (2017). Oil-palm yields in diversified plantations: Initial results from a biodiversity enrichment experiment in Sumatra, Indonesia. *Agriculture, Ecosystems & Environment*, *240*, 253–260. https://doi.org/10.1016/j.agee.2017.02.026

Gonzalez-Rivero, M., Beijbom, O., Rodriguez-Ramirez, A., Bryant, D. E. P., Ganase, A., Gonzalez-Marrero, Y., Herrera-Reveles, A., Kennedy, E. V., Kim, C. J. S., Lopez-Marcano, S., Markey, K., Neal, B. P., Osborne, K., Reyes-Nivia, C., Sampayo,

E. M., Stolberg, K., Taylor, A., Vercelloni, J., Wyatt, M., & Hoegh-Guldberg, O. (2020). Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. *Remote Sensing*, *12*(3), Article 489. https://doi.org/10.3390/rs12030489

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348–361. https://doi.org/10.3102/1076998619832248

Ismail, A., & Mamat, M. N. (2002). The optimal age of oil palm replanting. *Oil Palm Industry Economic Journal*, *2*(1), 11–18.

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation. *Remote Sensing*, *13*(21), Article 4405. https://doi.org/10.3390/rs13214405

Jelsma, I., Woittiez, L. S., Ollivier, J., & Dharmawan, A. H. (2019). Do wealthy farmers implement better agricultural practices? An assessment of implementation of Good Agricultural Practices among different types of independent oil palm smallholders in Riau, Indonesia. *Agricultural Systems*, *170*, 63–76. https://doi.org/10.1016/j.agsy.2018.11.004

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), Article 832. https://doi.org/10.3390/agriculture11090832

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. *Advances in Neural Information Processing Systems*, *30*, 3149–3157.

Legros, S., Mialet-Serra, I., Caliman, J. P., Siregar, F. A., Clément-Vidal, A., Fabre, D., & Dingkuhn, M. (2009). Phenology, growth and physiological adjustments of oil palm (*Elaeis guineensis*) to sink limitation induced by fruit pruning. *Annals of Botany*, *104*(6), 1183–1194. https://doi.org/10.1093/aob/mcp216

Machado, M. R., Karray, S., & de Sousa, I. T. (2019). LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In *2019 14th International Conference on Computer Science Education (ICCSE)* (pp. 1111–1116). IEEE. https://doi.org/10.1109/ICCSE.2019.8845529

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Morcillo, F., Cros, D., Billotte, N., Ngando-Ebongue, G. F., Domonhédo, H., Pizot, M., Cuéllar, T., Espéout, S., Dhouib, R., Bourgis, F., Claverol, S., Tranbarger, T. J., Nouy, B., & Arondel, V. (2013). Improving palm oil quality through identification and mapping of the lipase gene causing oil deterioration. *Nature Communications*, *4*(1), Article 2160. https://doi.org/10.1038/ncomms3160

Murphy, K. P. (2018). *Machine learning: A probabilistic perspective (adaptive computation and machine learning series)*. The MIT Press.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, *86*(4), 554–565. https://doi.org/10.1016/S0034-4257(03)00132-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Puttinaovarat, S., & Horkaew, P. (2019). Deep and machine learnings of remotely sensed imagery and its multi-band visual features for detecting oil palm plantation. *Earth Science Informatics*, *12*(4), 429–446. https://doi.org/10.1007/s12145-019-00387-y

Rahman, S. A. Z., Mitra, K. C., & Islam, S. M. (2018). Soil classification using machine learning methods and crop suggestion based on soil series. In *2018 21st International Conference of Computer and Information Technology (ICCIT)* (pp. 1–4). IEEE. https://doi.org/10.1109/ICCITECHN.2018.8631943

Raksaseri, K. (2023, February 5). *Thailand: Firm on protecting the palm oil sector*. https://www.reportingasean.net/thailand-firm-protecting-palm-oil-sector/

Sae-Tang, S. (2020). Estimation of oil content in oil palm fresh fruit bunch by its surface color. In *2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–5). IEEE. https://doi.org/10.1109/iSAI-NLP51646.2020.9376834

Samat, A., Li, E., Wang, W., Liu, S., Lin, C., & Abuduwaili, J. (2020). Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles. *Remote Sensing*, *12*(12), Article 1973. https://doi.org/10.3390/rs12121973

Sharma, R., Ghosh, A., & Joshi, P. K. (2013). Decision tree approach for classification of remotely sensed satellite data using open source support. *Journal of Earth System Science*, *122*, 1237–1247. https://doi.org/10.1007/s12040-013-0339-2

Shi, X., Cheng, Y., & Xue, D. (2019). Classification algorithm of urban point cloud data based on LightGBM. *IOP Conference Series: Materials Science and Engineering*, *631*(5), Article 052041. https://doi.org/10.1088/1757-899X/631/5/052041

Sinambela, R., Mandang, T., Subrata, I., & Hermawan, W. (2020). A ripeness study of oil palm fresh fruit at the bunch different positions. *Jurnal Keteknikan Pertanian*, *8*(1), 9–14. https://doi.org/10.19028/jtep.08.1.9-14

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919%2Fj.issn.1002-0829.215044

Suppalakpanya, K., Nikhom, R., Booranawong, T., & Booranawong, A. (2019). Forecasting oil palm and crude palm oil data in Thailand using exponential time-series methods. *Engineering & Applied Science Research*, *46*(1), 44–55. https://doi.org/10.14456/easr.2019.6

Swamynathan, M. (2017). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*.

Treerutkuarkul, A. (2021, January 11). *Making palm oil more sustainable*. https://www.bangkokpost.com/business/general/2048875/making-palm-oil-more-sustainable

Uning, R., Latif, M. T., Othman, M., Juneng, L., Mohd Hanif, N., Nadzir, M. S. M., Abdul Maulud, K. N., Jaafar, W. S. W. M., Said, N. F. S., Ahamad, F., & Takriff, M. S. (2020). A review of Southeast Asian oil palm and Its CO_2 fluxes. *Sustainability*, *12*(12), Article 5077. https://doi.org/10.3390/su12125077

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. *ACM Computing Surveys (CSUR)*, *53*(2), 1–33. https://doi.org/10.1145/3377454

Wangrakdiskul, U., & Yodpijit, N. (2015). Trends analysis and future of sustainable palm oil in Thailand. *Applied Science and Engineering Progress*, *8*(1), 21–32. https://doi.org/10.14416/j.ijast.2015.01.001

Worachairungreung, M., Thanakunwutthiro, K., & Kulpanich, N. (2023). A study on oil palm classification for Ranong province using data fusion and machine learning algorithms. *Geographia Technica*, *18*(1), 161–176. https://doi.org/10.21163/GT$_$2023.181.12

Xia, X., Pagano, A., Macovei, A., Padula, G., Balestrazzi, A., & Hołubowicz, R. (2024). Magnetic field treatment on horticultural and agricultural crops: Its benefits and challenges. *Folia Horticulturae*, *36*(1), 67–80. https://doi.org/10.2478/fhort-2024-0004

Zhong, Y., Liu, S., Luo, J., & Hong, L. J. (2022). Speeding up Paulson's procedure for large-scale problems using parallel computing. *INFORMS Journal on Computing*, *34*(1), 586–606. https://doi.org/10.1287/ijoc.2020.1054