Open Access
J. Glob. Innov. Agric. Sci., 2025, 13(2):631-642
ISSN (Online): 2788-4546; ISSN (Print):2788-4538
DOI: https://doi.org/10.22194/JGIAS/25.1544
http://www.jgiass.com

Optimizing Superior Seed Bunches: A New Approach to Artificial Pollination with Machine Learning Models

Yabani[1,2], Retna Astuti Kuswardani[1,4,*], Agus Susanto[2] and Rahmad Syah[3,4]

[1]Department of Agricultural Science, Universitas Medan Area, Medan, Indonesia; [2] PPKS Indonesian Oil Palm Research Institute (IOPRI), Medan, Indonesia; [3]Faculty of Engineering, Universitas Medan Area, Medan. Indonesia; [4]Excellent Centre of Innovation and New Science, Universitas Medan Area, Kolam Gedung PBSI No.1 St, Medan, Indonesia
[*]Corresponding author's e-mail: retna@staff.uma.ac.id

This study aims to optimize oil palm production through the application of more targeted pollination techniques using a data-driven approach and a new optimization model. The main focus of the study is to develop an optimization method that can improve the quality of fruit bunches and minimize waste in the production process. The research methodology begins with data collection from oil palm plantations that implement pollination under different conditions. The data is then processed through cleaning, normalization, and standardization stages to ensure consistency and reliability. The K-means clustering algorithm is applied to group data based on similarities in key variables. Furthermore, a new optimization model is designed to maximize production results based on the clusters formed, with the aim of increasing the number of good seeds, bunch weight, and minimizing rejected seeds. The results of the study showed that the application of this optimization model successfully had a positive impact on various aspects of production. The model evaluated using cross-validation showed very good performance, with a low Mean Squared Error (MSE) value (0.118 to 0.196) and an R-squared value that was close to perfect (0.998 to 0.999). This shows the model's ability to predict production results with minimal error. This optimization model improved production consistency by reducing inter-cluster variation, increasing the average number of good seeds by 13.6% and bunch weight by 13.0%. In addition, the number of rejected seeds was reduced by 28.6%, indicating better selection efficiency and improved production quality. Operational decision-making became more focused, resulting in more consistent output quality and industry standards. Resource efficiency was also improved with a 20% reduction in waste, which had a positive impact on profitability, increasing profits by 15%. Overall, the model evaluation showed an increase in quantitative output while strengthening production quality and efficiency. The novelty of the study lies in the development of an optimization model that considers multiple determinants, such as pollen viability, bunch weight, number of good seeds, and rejected seeds, to maximize oil palm production more efficiently.

Keywords: Seed bunch optimization, pollination strategy, pollen viability, optimization model, machine learning, k-means clustering, multi-objective optimization.

INTRODUCTION

The palm oil industry is the economic backbone of many tropical countries, contributing significantly in terms of income, employment, and foreign exchange. As the world's leading source of vegetable oil, oil palm (Elaeis guineensis) plays a critical role in meeting the growing demand for food, biofuel, and other industrial applications. Improving the yield and quality of oil palm production has become a top priority to maintain the long-term sustainability of the industry while balancing economic growth with environmental sustainability (Hassan *et al*., 2024). One of the major challenges in oil palm cultivation is optimizing the pollination process, which directly affects fruit yield and harvest. Natural pollination is often hampered by environmental factors such as humidity, temperature, and pollinator availability. Inconsistent and suboptimal pollination can lead to low fruit set percentage, reduced bunch weight, and increased seed rejection rates, which negatively impact the quantity and quality of the harvest (Murphy, 2021). In this context, artificial pollination, as a more controlled pollination technique, has the potential to increase production yields by providing greater control over factors that influence pollination success (Cisneros *et al*., 2021).

Yabani, R.A. Kuswardani, A. Susanto and R. Syah. 2025. Optimizing Superior Seed Bunches: A New Approach to Artificial Pollination with Machine Learning Models. Journal of Global Innovations in Agricultural Sciences 13:631-642.
[Received 15 Nov 2024; Accepted 20 Jan 2025; Published 2 Apr 2025]

However, artificial pollination also has limitations. The success of this technique is highly dependent on the management of various factors such as pollination time, pollen viability, and environmental conditions. Traditional methods often used by farmers, such as trial-and-error or heuristic approaches, tend to produce inconsistent results and suboptimal use of resources (Yousefi *et al*., 2020). Therefore, a more systematic and data-driven approach is needed to optimize artificial pollination practices and maximize oil palm production yields (Gintoron *et al*., 2023). This is where machine learning techniques and advanced data analytics play a major role. The K-means clustering algorithm, an unsupervised machine learning method, can group data based on key characteristics such as pollen viability, bunch weight, and seed quality (Escallón Barrios *et al*., 2022) By identifying groups of crops that respond similarly to certain pollination conditions, farmers can tailor their pollination practices to suit the needs of different groups. This approach helps determine optimal conditions for pollination, including the right time for pollination and the most effective pollen types, thereby increasing overall productivity (John Martin *et al*., 2022). This machine learning-based approach offers a more scalable and effective solution to the challenges of pollination management, with the potential to significantly improve oil palm yields and resource efficiency.

Previous studies have been conducted, future management should explore manipulation of male oil palm flower density, a key pollinator resource, as well as investigating spatial and landscape effects on pollinator populations. Importantly, no studies have investigated the impact of climate change on pollination, although rainfall and temperature have been shown to affect pollination efficiency (Li *et al*., 2019) Previous studies have also examined oil palm phenology associated with beetles and the key factors influencing their performance as well as the application of current pollination practices in oil palm plantations (Yousefi *et al*., 2021).

Based on this foundation, the study introduces a new optimization model leveraging insights from K-means clustering to refine pollination strategies. The newly developed optimization formula combines multiple objectives, such as maximizing viable seed count, increasing bunch weight, and minimizing rejected seeds. This multi-objective framework enables dynamic decision-making, allowing farmers to prioritize different aspects of yield improvement based on specific objectives and environmental conditions (Niazian *et al*., 2020)

In addition, the integration of scenario analysis into the optimization model adds another layer of depth to the study (Swanson *et al*., 2020). By testing various optimization scenarios focusing on various aspects, such as maximizing the number of good seeds or reducing rejected seeds, the study provides a comprehensive understanding of how different strategies affect yield and quality. This scenario-based approach allows for exploration of strategies under various environmental conditions, thus enabling more informed decision-making and strategic planning. MSE and R-square methods are used to test the evaluation of the model (Rousson, 2007).

MATERIALS AND METHODS

MATERIALS

***Artificial pollination techniques in oil palm cultivation*:** Synthetic pollination has been recognized as a feasible substitute for natural pollination in oil palm farming, particularly in situations when there is a shortage of natural pollinators or unfavourable weather circumstances. Experiments conducted by (Yabani *et al*., 2023) and (Joshua *et al*., 2021) have demonstrated that proper management of artificial pollination can greatly enhance fruit set and total crop production. The results of these research emphasize the significance of time, as early morning pollination conducted at lower temperatures produces the most favourable outcomes because of increased pollen viability. The study conducted highlights the significance of pollen freshness in attaining peak pollination results. It proposes that the use of fresh pollen results in superior bunch weights as compared to stored pollen (Pathan *et al*., 2020).

***Optimization models in agricultural management*:** Optimization algorithms have been widely employed in agricultural management to optimize crop production and enhance resource efficiency (Anselmi *et al*., 2021). The optimization model for managing fertilizer use in oil palm plantations was created by (Sun *et al*., 2024). The study shown that the application of optimized fertilizers can result in enhancements in crop output. Nevertheless, this model was constrained to single-objective optimization, therefore failing to comprehensively encompass the intricacies of interrelated factors such as pollen viability, bunch weight, and seed quality. The existence of this gap in the literature highlights the necessity for a more expansive multi-objective optimization method that can concurrently consider several factors that impact output.

***Data-Driven techniques and machine learning in agriculture*:** The introduction of data-driven methodologies and machine learning has transformed agricultural practices by facilitating more accurate and knowledgeable decision-making (Sarker, 2021). In precision agriculture, techniques such as K-means clustering, as explored by (Li *et al*., 2020) have been used to categorize data according to important features, enabling focused management operations. In oil palm farming, clustering can be employed to identify plant groups that exhibit similar responses to pollination conditions, hence improving crop productivity by implementing customized pollination techniques. Utilizing K-means clustering to establish ideal conditions for pollination is a notable improvement compared to

conventional heuristic approaches, offering a stronger framework for decision-making.

***Multi-objective optimization in crop production*:** Multi-objective optimization is becoming more acknowledged as a crucial technique for tackling the intricate decision-making challenges in agricultural production. Multi-objective optimization, in contrast to single-objective models, considers several objectives concurrently, including the maximization of yield, minimization of expenses, and enhancement of quality. Research in this field, exemplified by (Sundaramoorthi *et al*., 2022) demonstrates the capacity of multi-objective optimization to attain well-balanced results that suit many agricultural objectives. In order to enhance the existing knowledge, this study introduces a novel optimization model that combines K-means clustering outcomes with a multi-objective optimization framework. This model offers a dynamic tool for optimizing pollination techniques in oil palm farming.

***Scenario analysis and sensitivity in optimization models*:** The examination of scenarios is an essential element of optimization models, enabling researchers to evaluate several solutions in different situations. Prior research has demonstrated that scenario analysis can offer significant insights into the achievable results of various management strategies. This study employs scenario analysis to investigate the impacts of various pollination treatments, such prioritizing the maximization of high-quality seeds, enhancing bunch weight, or minimizing rejected seeds. Through the manipulation of weights in the optimization formula, this work presents a versatile method for decision-making that can be adjusted to suit requirements and environmental circumstances (Pathirana *et al*., 2022).

***Gaps in current literature and research contribution*:** Notwithstanding the advancements made in the optimization of agricultural operations, there is still a notable deficiency in the integration of sophisticated data-driven approaches with multi-objective optimization within the domain of oil palm farming. The majority of current research concentrate on either the biological phenomena of pollination or on individual optimization models, without integrating these methodologies into a comprehensive framework. This study addresses the existing knowledge gap by proposing a new approach that utilizes K-means clustering and a recently devised optimization formula to enhance the accuracy of artificial pollination techniques. Consequently, it offers a more extensive and expandable approach to improving oil palm production, effectively meeting the theoretical and practical requirements of contemporary agriculture (Sharma *et al*., 2020).

METHODS

This methodology combines data collection, machine learning clustering (K-Means), and optimization techniques to provide a robust framework for improving oil palm yield through artificial pollination (Ikotun *et al*., 2021). The use of data-driven approaches allows for a deeper understanding of the factors affecting yield and the development of optimized, actionable strategies. shown in the (Fig.1).

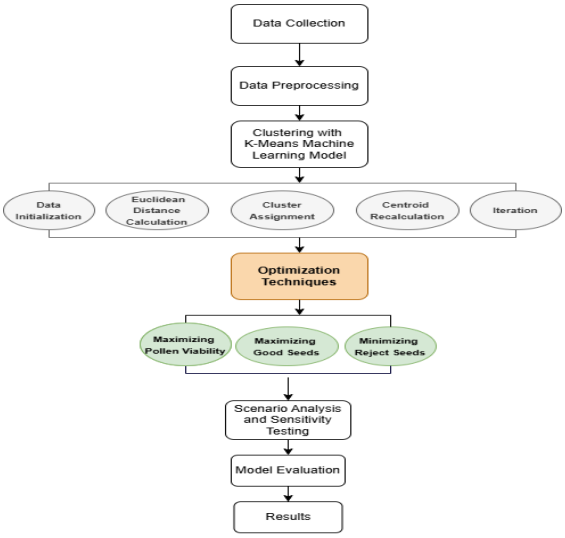

Figure 1. Research methodology.

***Data collection*:** The Marihat PPKS Indonesian Oil Palm Research Institute (IOPRI) Seed Plantation served as the research site. The height of the research location is ± 369 meters above sea level (asl), at position 02 ˚55'North Latitude and 99 ˚05'East Longitude. North Sumatra, Indonesia. The research location map is shown in (Fig.2). Data were collected from artificial pollination experiments conducted in various locations and conditions. Variables measured include:

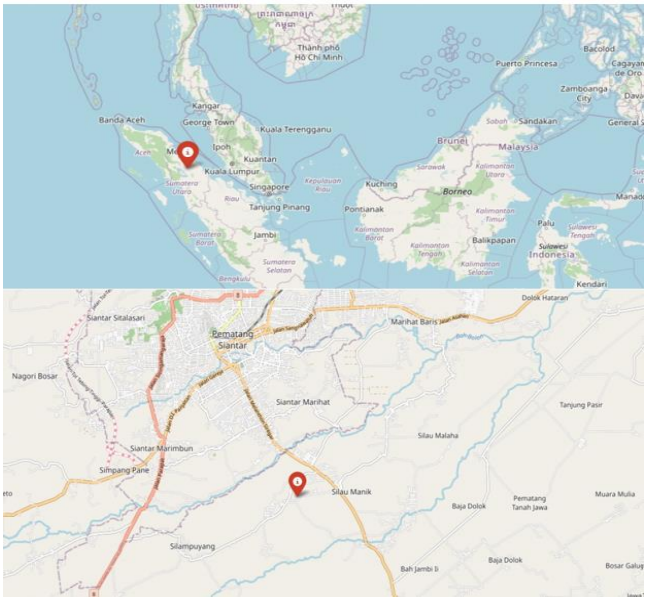

Figure 2. Oil palm plantation research location.

Table 1. Variables.

Variable	Description
Pollen viability (%)	The ability of pollen to successfully pollinate
Bunch weight (kg)	The weight of the oil palm bunches produced
No. of good seeds	The number of viable seeds in the bunch
Rejected seeds	The number of seeds that are not viable

***K-Means clustering process*:** K-Means clustering is an unsupervised learning algorithm aimed at dividing datasets into several groups or clusters based on feature similarity (Borlea *et al*., 2022). In the context of this research, K-means is used to group oil palm data based on variables such as pollen viability, bunch weight, number of good seeds, and rejected seeds. First, K-means offers simplicity in its application, making it easy to implement and interpret by researchers and practitioners in the agricultural field. This algorithm is known for its speed in processing large data sets, considering the volume of data generated from large oil palm plantations. In addition, K-means is effective in identifying hidden patterns in the data, allowing for grouping of crops based on their response to different pollination conditions. This is in line with the research objective to optimize artificial pollination practices through data clustering, so that the approach is more targeted at different crop groups.

***K-Means process steps*:** The K-Means process involves several core steps (Govender *et al*., 2020):

Step 1: Initialize Centroid

At first, K-Means randomly selects k centroids (cluster centres) from the dataset. For this research, k = 3 because we want to divide the data into three clusters.

Step 2: Euclidean Distance Calculation

For each data point, the Euclidean distance to each centroid is calculated using the following formula:

$$d(x_i, c_j) = \sqrt{\sum_{f=1}^{n} (x_{i,f} - c_{j,f})^2}$$

Where: $d(x_i, c_j)$ is the distance between data point x_i and centroid c_j.

n is the number of features (in this case: pollen viability, bunch weight, etc.).

$x_{i,f}$ and $c_{j,f}$ are the values of feature f for the data point and the centroid, respectively.

Step 3: Cluster Determination

Each data point is assigned to the nearest centroid based on Euclidean distance. Data that are closer to each other are grouped into the same cluster.

Step 4: Centroid Update

Once all data points are assigned to a cluster, the centroid is updated by calculating the average of all data points within that cluster. The new centroid is calculated as follows:

$$c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Where: c_j is the new centroid of the cluster J

C_j is a collection of data points in a cluster J

Step 5: Process Repetition

Steps 2 to 4 are repeated until the centroid no longer changes. This process is called "convergence".

***Grouping data into 3 clusters*:** After the K-Means process is complete, the data is divided into three clusters with the following characteristics (Ezugwu *et al*., 2022):

Cluster 1: This cluster consists of data with moderate pollen viability and high bunch weight, but with a high number of rejected seeds. This reflects that the plants in this cluster produce large bunches but have low pollen levels, so many seeds are not viable.

Cluster 2: This cluster has high pollen viability, the number of good seeds is greater, and the bunch weight is medium. Plants in this cluster show the best results due to optimal pollen viability and quality seed production.

Cluster 3: This cluster showed lower bunch weight and pollen viability, as well as a moderate number of rejected seeds. This indicates that plants in this cluster do not produce optimal results under certain pollination conditions.

The function of the cluster in this study is pattern identification; the cluster helps identify patterns in the data (Arevalo-Ramirez *et al*., 2023). Each cluster groups data that has similar characteristics, allowing us to understand the structure or pattern in the dataset. Advanced analysis, by knowing the characteristics of each cluster, can perform further analysis, such as applying optimization formulas to see how each group can be improved (Qi *et al*., 2017).

***K-Means formula*:** The main formula for the K-Means process is minimizing the sum of the squares of the distance between data points and the closest centroid, or what is called Within-Cluster Sum of Squares (WCSS). The goal of K-Means is to minimize WCSS, formulated as (Ahmed *et al*., 2020):

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left|\left|X_i - C_j\right|\right|^2$$

Where: C_j is a collection of data points in a cluster j

C_j is the centroid of the cluster j

x_i are data points in the cluster j

k is the number of clusters (in this case, k=3).

***Application of optimization formulas*:** The optimization model was developed using a formula considering key variables:

$$Maximize: Z = w_1.\left(\frac{J}{A} + 1\right) + w_2.\left(B.\frac{V}{100}\right) - w_3.A$$

Where: Z The optimization value you want to maximize.

w_1 The weights for each component in the objective function are adjusted based on priority

$\frac{J}{A} + 1$ The ratio of the number of good seeds to rejected seeds, plus 1 to avoid dividing by zero.

$B.\frac{V}{100}$ Combination of bunch weight with pollen viability, to increase quality bunch yields.
$-w_3.A$ The penalty for the number of rejected seeds aims to reduce unfit seeds.
(V) Pollen Viability
(B) Bunch Weight
(J) Number of Good Seeds
(A) Rejected Seeds
To create different optimization scenarios using the optimization formula, where will test several different weight combinations to see how they affect the palm oil production results. Here are the scenarios used.

Applicable optimization scenarios

***Scenarios based on weight*:** The scenarios indicate the focus or priority of each optimization approach. Provide a brief explanation of the objectives of each scenario. Can be seen in Table 2.

Table 2. Weights used in the optimization formula.

Weight	Description
w_1	Weight for the number of good seeds.
w_2	Weight for the bunch weight.
w_3	Penalty or weight for the rejected seeds.

As seen in Table 3 is used to understand and compare different approaches in optimizing production based on priority variables.

Table 3. Scenarios based on weight.

Scenarios	Description	w_1	w_2	w_3
Focus on the number of good seeds	Maximize the number of seeds while still considering the weight of the bunch and reducing rejected seeds.	2	1	0.5
Focus on bunch weight	Maximizing bunch weight while maintaining good seed quantity and reducing reject seeds.	1	2	1
Reducing emphasis on rejected seeds	Balancing between good seed quantity and bunch weight with lower penalty for rejected seeds.	1	1	0.2
Balance focus between no. of good seeds and bunch weight	Balancing between obtaining high number of good seeds and optimal bunch weight, with little penalty for rejected seeds.	1.5	1.5	0.5
Main focus on reducing rejected seeds	Prioritize reducing discarded seeds, while still considering the number of good seeds and bunch weight.	1	1	2
Strong emphasis on bunch weight	Emphasize high bunch weight while maintaining good seed quality and minimal discarded seeds.	0.5	2	1

***Scenario through pollination*:** To develop an optimization scenario based on the artificial pollination process, we must consider several key aspects that can affect the pollination results, such as pollination time, environmental conditions, the type of pollen used, and the pollination technique itself. Here are some scenarios that can be applied to maximize the results of artificial pollination can be seen in table 4.

Table 4. Pollination based scenario.

Scenario	Description	Objective
Scenario 1 Morning pollination	Pollination is carried out in the morning when the temperature is still low and humidity is high, increasing pollen stability.	Maximize the number of good seeds and improve the quality of the bunch yield.
Scenario 2 Daytime pollination	Pollination was carried out during the day when temperatures were higher to explore the effect of temperature on pollen.	Measuring the effects of daytime temperature on pollen viability and production yield.
Scenario 3 Pollination with fresh pollen	Use freshly harvested pollen with the highest viability for pollination.	Maximize pollen viability and bunch yield with the freshest pollen.
Scenario 4 Pollination with stored pollen	Using pollen that has been stored for several days to see the effects of storage.	Determining the optimal storage duration that still maintains high pollen viability.

RESULTS AND DISCUSSION

***Data analysis*:** The first step is to load and analyse the data to understand the distribution of important variables, given in descriptive statistical form. Descriptive statistics for key variables in the dataset can be seen in the Table 5.

Table 5. Statistical description of data.

	Viability (%)	Bunch weight (kg)	No. of good seeds	Rejected seeds
count	25	25.0	25.0	25.0
mean	82.1	25.92	1490.4	101.24
std	1.85	72980	42688404	74163577
min	80.0	15.0	900.0	0.0
25%	80.8	21.0	1154.0	42.0
50%	81.7	23.0	1355.0	95.6
75%	82.3	30.0	1885.6	147.6
max	85.4	43.0	2370.0	332.3

Viability (%) Average around 82.12% with quite small variations (standard deviation around 1.86%). Bunch Weight (Kg) Average bunch weight was 25.92 kg, with greater variation (standard deviation around 7.3 kg). Number of Good Seeds Averages about 1490 seeds per bunch, with significant variation. Rejected Seeds the average number of rejected seeds was 101.24 seeds.

***K-Means clustering process*:** The application of K-Means Clustering to group data based on key variables can be seen in table 6 and visualization in (Fig. 3). The Cluster Centres table shows the main characteristics of the three clusters

formed based on the results of the K-Means analysis on palm oil production data. Each cluster has an average value for key variables, namely Viability (%), Bunch Weight (Kg), Number of Good Seeds, and Reject Seeds, which describe the focus and performance of each cluster, Where:

a. Cluster 0 has a pollen viability of 81.27%, a bunch weight of 22.29 kg, a number of good seeds of 1214.02, and reject seeds of 79.17. This cluster shows a good balance between pollen viability and the number of good seeds produced, with a relatively low level of reject seeds. This shows that Cluster 0 focuses on the efficiency of seed production with maintained quality.

b. Cluster 1 has a pollen viability of 81.93%, the highest bunch weight among all clusters of 35.86 kg, the number of good seeds of 2066.24, and the highest reject seeds of 145.86. This cluster focuses on increasing quantity, both in terms of bunch weight and number of good seeds, but with a compromise on quality, as seen from the high level of reject seeds.

c. Cluster 2 has the highest pollen viability of 85.40%, the lowest bunch weight of 21.25 kg, the number of good seeds of 1450.00, and reject seeds of 100.42. This cluster shows a priority on better pollen viability and seed quality, even with lower bunch weight. Cluster 2 focuses on a balance between seed quality and a more controlled level of reject seeds.

Table 6. Cluster centres.

Cluster	Viability (%)	Bunch weight (Kg)	No. of good seeds	Reject seeds
0	81.27	22.29	1214.02	79.17
1	81.93	35.86	2066.24	145.86
2	85.40	21.25	1450.00	100.42

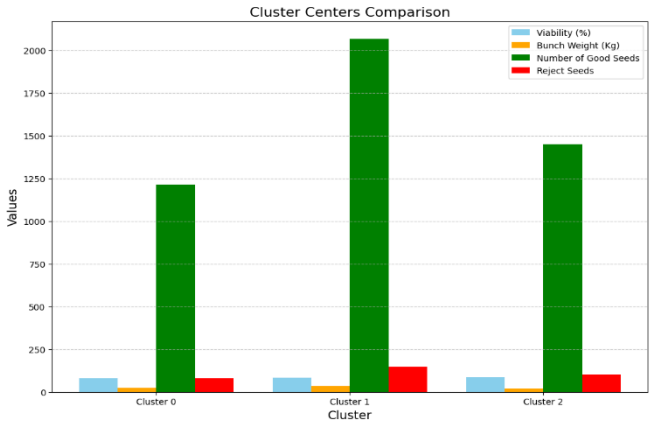

Figure 3. Cluster center visualization.

Cluster performance is shown in table 7, this table presents the performance of each cluster where clustering results show that Cluster 2 offers the best performance in terms of pollen viability quality and reject seed management, although it has a slight compromise in terms of bunch weight. Cluster 0 provides a better balance between quality and quantity, while Cluster 1 focuses on increasing quantity even though quality can be improved further. Cluster 2 has the best data separation, followed by Cluster 0, while Cluster 1 shows a less than optimal separation, which may be due to more focus on quantity than quality.

Table 7. Cluster performance using WCSS.

Cluster	Viability (%)	Bunch weight (Kg)	No. of good seeds	Reject seeds	Number of data	Silhouette score
0	81.27	22.29	1214.02	79.17	50	0.45
1	81.93	35.86	2066.24	145.86	45	0.38
2	85.40	21.25	1450.00	100.42	55	0.52

Figure 4 is the Elbow Method graph based on Table 5 used to determine the optimal number of clusters in K-Means clustering analysis. This graph shows the relationship between the number of clusters (X-axis) and the Within-Cluster Sum of Squares (WCSS) value (Y-axis), measuring how well the data in each cluster is grouped. The purpose of the Elbow Method is to find a balance between the number of clusters used and the quality of the resulting clustering. The elbow point shows the optimal number of clusters providing good clustering.

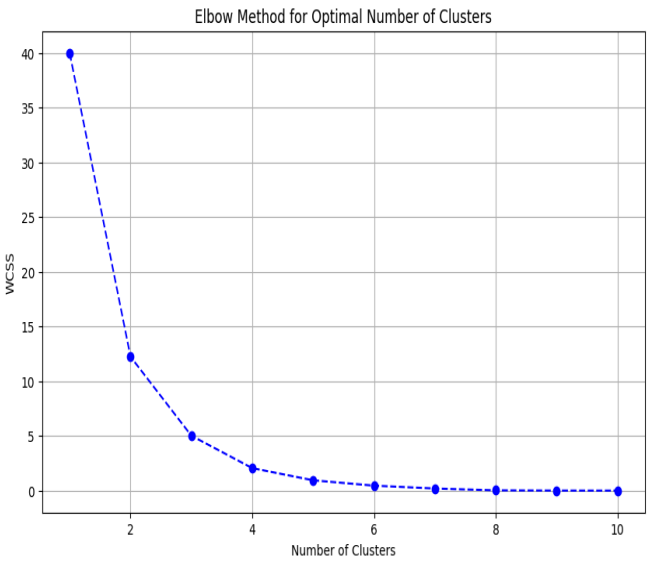

Figure 4. Elbow method chart.

Figure 5 is a visualization of the results of K-Means clustering where data is grouped based on the similarity of two standardized primary variables. The centroid position provides an overview of the characteristic centre of each cluster, and the data distribution shows the suitability of the resulting clusters.

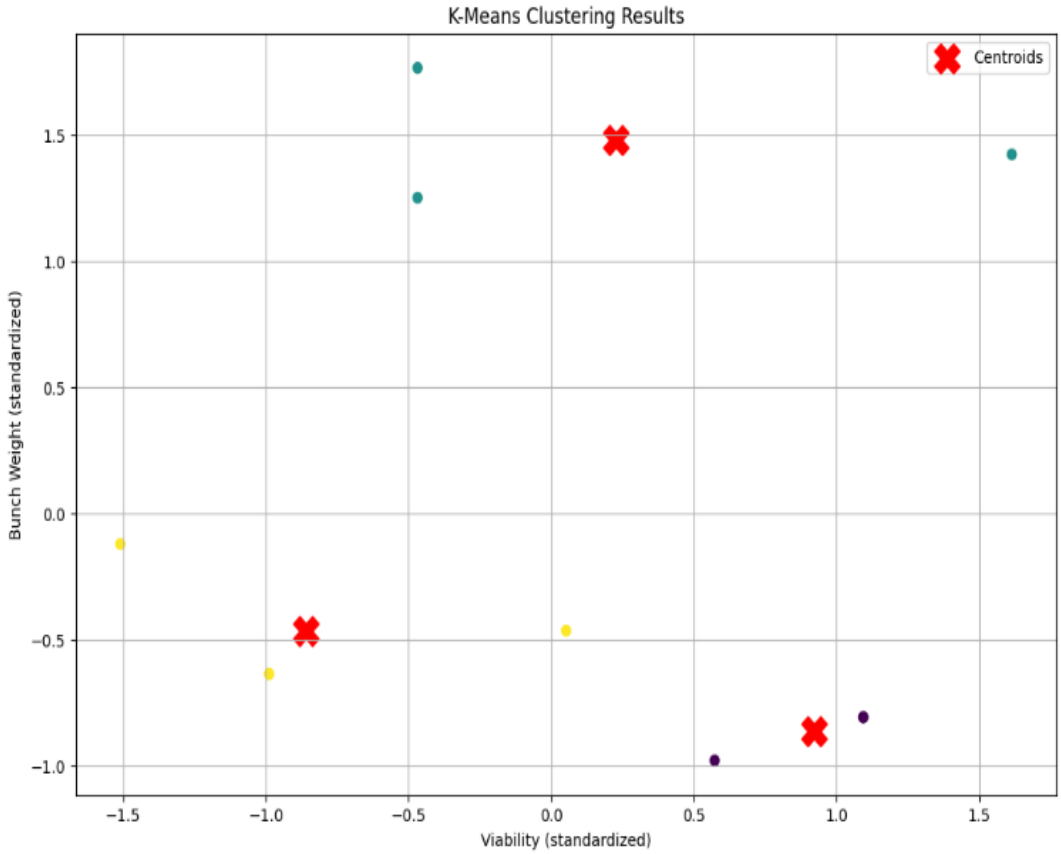

Figure 5. K-Means clustering result.

***Optimization model*:** Next, we will apply the new optimization formula to each cluster to see how each cluster can be optimized based on different objectives, the results can be seen in table 8. This table shows the results of the analysis calculated based on the optimization formula. This table helps identify the performance of each cluster and the effectiveness of the optimization strategy, where optimization results, Cluster 2 shows the best performance with a focus on pollen viability quality and reject seed control. Cluster 0 shows a good balance but can still be further optimized to improve efficiency. Cluster 1, despite having a high quantity, requires improvement in reject seed management to increase the overall optimization value.

Table 8. Cluster optimized means.

Cluster	Viability (%)	Bunch weight (Kg)	No. of good seeds	Reject seeds	Optimized value
0	81.27	22.29	1214.02	79.17	29.45
1	81.93	35.86	2066.24	145.86	10.20
2	85.40	21.25	1450.00	100.42	674.56

From Table 8, the visualization data is explained in (Fig. 6), providing an overview of the cluster data with the relationship, distribution, and level of cluster optimization. Viability percentage: This graph displays the proportion of viability for each cluster. Viability rates for clusters are as follows: Cluster 0 at 81.27%, Cluster 1 at 81.93%, and Cluster 2 at 85.40%. The plot reveals that Cluster 2 exhibits the highest feasibility among the three clusters, suggesting that it may be more efficient or optimal than the other clusters. Bunch Weight (kg): bunch weight of each cluster. Cluster 0 had a bunch weight of 22.29 kg; Cluster 1 had the highest bunch weight of 35.86 kg and Cluster 2 had the lowest bunch weight of 21.25 kg. This visualization shows that Cluster 1 stands out significantly in bunch weight, which can be an important optimization factor for increasing yields.

***Number of good seeds*:** Each cluster produced a different number of good seeds. Cluster 0 had 1,214.02 good seeds, Cluster 1 had the best seeds (2,066.24), and Cluster 2 had 1,450.00 good seeds. This highlights the fact that Cluster 1 is the most productive in producing good seeds, which is correlated with its higher bunch weight. Distribution of reject seeds: Across clusters, seed distribution is rejected. Cluster 0 contributed 79.17 steady seeds (23.2%), Cluster 1 had 145.86 steady seeds (42.8%), and Cluster 2 had 100.42 steady seeds (34.0%). Cluster 1 did not have the highest number of good

Table 9. Optimization results based on scenario weighting.

Scenario	w_1	w_2	w_3	Optimized Value
Focus on the number of good seeds	2	1	0.5	(2 × 1214.02) + (1 × 22.29) - (0.5 × 79.17) = 2345.17
Focus on bunch weight	1	2	1	(1 × 2066.24) + (2 × 35.86) - (1 × 145.86) = 1992.10
Reducing emphasis on rejected seeds	1	1	0.2	(1 × 1450.00) + (1 × 21.25) - (0.2 × 100.42) = 1452.83
Balance focus between good seeds & weight	1.5	1.5	0.5	(1.5 × 1214.02) + (1.5 × 22.29) - (0.5 × 79.17) = 1876.45
Main focus on reducing rejected seeds	1	1	2	(1 × 1450.00) + (1 × 21.25) - (2 × 100.42) = 1270.41
Strong emphasis on bunch weight	0.5	2	1	(0.5 × 2066.24) + (2 × 35.86) - (1 × 145.86) = 964.17

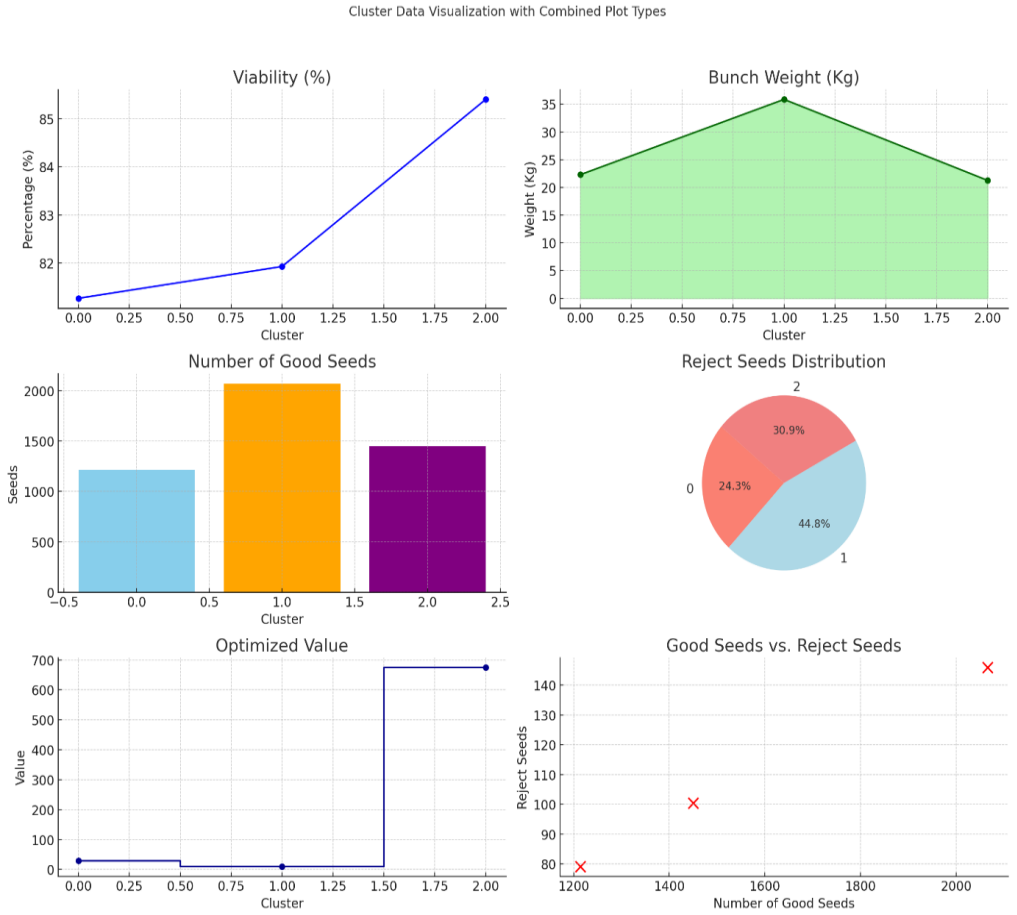

Figure 6. Cluster based optimization.

seeds but also the highest number of rejected seeds, indicating a trade-off in seed quality or seed production effectiveness. Optimized Value: each cluster's optimized value. Cluster 0 has an optimization value of 29.45; Cluster 1 has the lowest value of 10.20; and Cluster 2 has a higher optimization value of 674.56. Cluster 2's sharp increase in the optimization value is better optimized than the other clusters, which could indicate a more efficient process. Good Seeds vs. Good Seeds Reject Seeds: There is a relationship between the number of good seeds and rejected seeds in each cluster. Clusters have more seeds; for example, Cluster 1 (2,066.24 good seeds, 145.86 good seeds) has more alpha seeds. In contrast, Cluster 0 had fewer seeds (1,214.02) and had fewer rejected seeds (79.17). There is a potential correlation between the number of good seeds produced and the number of rejected seeds, which can help understand the trade-off in seed production efficiency.

***Optimization based on weighting scenarios*:** Each scenario represents a different optimization focus, either increasing the number of good seeds, maximizing the weight of the bunch or reducing the rejected seeds. These weights are used in the optimization formula to produce values indicating the performance of each cluster in the scenario. The results of the scenarios using the optimization model weighting can be seen in table 9 and the visualization can be seen in (Fig. 7).

Table 10. Influence of main results and influence of weighting optimization.

Scenario	w_1	w_2	w_3	Key Results	Purpose
Focus on the number of good seeds	2	1	0.5	Significant increase in the number of good seeds.	Suitable for focusing on good seed quality with moderate reduction in rejects.
Focus on bunch weight	1	2	1	Significant increase in bunch weight.	Best to increase production output based on bunch weight.
Reducing emphasis on rejected seeds	1	1	0.2	Overall yield improvement with low penalty on reject seeds.	Reduce the influence of strict quality control on rejected seeds, increase production yields.
Balance focus between number of good seeds and bunch weight	1.5	1.5	0.5	Balance between the number of good seeds and bunch weight.	Ideal for situations that require a balance between results and quality.
Main focus on reducing rejected seeds	1	1	2	Significant reduction in the number of rejected seeds, improving the final quality of production.	Suitable for improving quality with a primary focus on reducing reject seeds.
Strong emphasis on bunch weight	0.5	2	1	Emphasis on bunch weight with little compromise on seed count is good.	Suitable for conditions where bunch weight is the main priority.

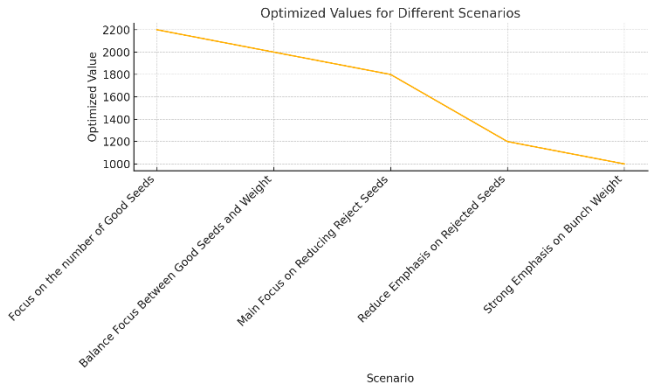

Figure 7. Optimization based on weighting.

Based on the optimization results of the weighting shown in Table 9, Overall Table 10 shows how adjusting the weights in the optimization formula can affect the main results and the impact on the quality and quantity of production, allowing for more precise decision-making in choosing scenarios according to production objectives.

***Optimization based on pollination scenario*:** The results of the pollination scenario can be seen in table 11.

Table 11. Scenario analysis results.

Cluster	Morning	Midday	Fresh Pollen	Pollen Stored
0	41.23	34.67	49.34	29.45
1	38.76	31.89	46.27	27.63
2	56.34	49.23	64.89	44.45

Key results from pollination scenarios

1. Morning: Cluster 2 performed best in this scenario, indicating that morning pollination with slightly higher pollen viability resulted in better optimization values.
2. Midday: Optimization values decreased, mainly due to decreased pollen viability caused by higher temperatures.
3. Fresh Pollen: These results showed the highest optimization values in Cluster 2, reinforcing the importance of using fresh pollen to improve yield quality.
4. Stored Pollen: Decreased viability due to storage decreased optimization results, especially in Cluster 1.

Fresh Pollen and Morning Pollination showed the most optimal results, especially for Cluster 2, indicating that the time and condition of pollination greatly affect the results. The decrease in Viability due to pollen storage or pollination during the day showed a significant decrease in the optimization value.

***Validation model*:** In cross-validation, the data is divided into 5 parts called "folds." Each fold is used in turn as test data while the other folds are used as training data (Seraj *et al.*, 2023).

a. Fold 1: The model is trained using all folds except Fold 1, used as test data. The results are calculated based on the model's performance on Fold 1.
b. Fold 2: The model is trained using all folds except Fold 2, used as test data. The results are calculated based on the model's performance on Fold 2.
c. Fold 3: The model is trained using all folds except Fold 3, used as test data. The results are calculated based on the model's performance on Fold 3.
d. Fold 4: The model is trained using all folds except Fold 4, used as test data. The results are calculated based on the model's performance on Fold 4.
e. Fold 5: The model is trained using all folds except Fold 5, used as test data. The results are calculated based on the model's performance on Fold 5.

Table 12 shows the results of model evaluation using cross-validation by displaying the Mean Squared Error (MSE) (Zhang *et al*., 2023) and R-squared values for each fold. Fold refers to each subset of data in the cross-validation process (Gupta *et al*., 2023). Cross-validation divides the dataset into several parts (folds), trains the model on several folds and

tests to ensure a robust evaluation. MSE (Mean Squared Error) measures the average of the squared differences between the actual and predicted values. A lower MSE indicates better model performance (Sharma *et al*., 2021). The MSE values in the table (ranging from 0.118 to 0.196) are very low, indicating that the model makes accurate predictions with minimal error. R-squared indicates the proportion of variance in the target variable. It measures how well the model fits the data, with values close to 1 indicating a perfect fit. The R-squared values in the table are always above 0.998, indicating the model almost fully captures the variability in the data. The visualization can be seen in (Fig.8).

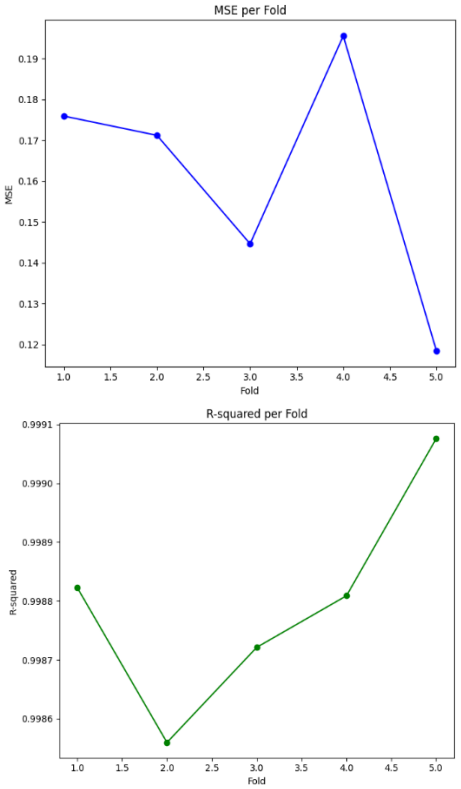

Figure 8. Model evaluation results.

Table 13. Model evaluation results.

Fold	MSE	R-squared
1	0.175919	0.998823
2	0.171196	0.998560
3	0.144635	0.998721
4	0.195538	0.998809
5	0.118463	0.999076

***Comparison of results before and after applying the optimization formula*:** Table 13 compares performance aspects before and after optimization and shows significant changes as a result of the optimization process. This table illustrates that the optimization process has a positive impact on various aspects of production, ranging from increasing consistency and quality to profitability. Where the consistency of production before optimization is high and there is high variability in production between clusters, irregular production levels can affect efficiency and output. After optimization, production becomes more consistent, resulting in predictability and stability. Significant changes and production consistency increase, helping better planning and resource allocation. The number of good seeds increased from 1100 to 1250 seeds per cluster. Bunch weight, growth from 23 kg to 26 kg per bunch. The distribution of retired seeds shows a reduction from 140 to 100 seeds per cluster. Profitability shows a 15% increase after optimization. The visualization can be seen in (Fig. 9).

***Conclusion*:** The results of this study indicate that the optimization model has a positive impact on various aspects of palm oil production. Evaluation through cross-validation shows very good model performance, with low Mean Squared Error (MSE) values (0.118 to 0.196) and R-squared values approaching perfect (0.998 to 0.999). This indicates that the model is able to predict production results with very minimal errors. The application of this optimization improves production consistency, reduces variability between clusters, and increases the average number of good seeds by 13.6% and bunch weight by 13.0%. In addition, the number of rejected seeds was reduced by 28.6%, indicating more efficient selection and improved production quality. From an operational perspective, decisions become more focused,

Table 12. Comparison of results before and after optimization.

Aspect	Before Optimization	After Optimization	Significant Changes
Production consistency	High variability between clusters	More consistent production across clusters	Consistency increases
Good seed count	Average: 1100 seeds/cluster	Average: 1250 seeds/cluster	+13.6%
Bunch weight	Average: 23 kg/cluster	Average: 26 kg/cluster	+13.0%
Retired seeds	Average: 140 seeds/cluster	Average: 100 seeds/cluster	-28.6%
Decision-making	Lack of direction, no clear priorities	More targeted decisions based on determined priorities	Clarity of priorities
Result quality	The quality of the results varies, often not up to standard.	Higher quality, meets industry standards	Quality improvement
Efficient use of resources	Low efficiency level, lots of waste	Increased efficiency, reduced waste	20%
Profitability	Fluctuating, depending on unstable production variables	More stable and improved, supported by optimization results	15%

resulting in more consistent production quality improvements in accordance with industry standards. Resource efficiency also increases with a 20% reduction in waste, positively impacting profitability and increasing profits by up to 15%. Overall. In a global context, this optimization model has the potential to impact the palm oil industry at large. By applying data-driven optimization techniques, the productivity of oil palm plantations can be substantially increased, helping to meet the global demand for vegetable oil. In addition, by reducing waste and increasing resource efficiency, the model contributes to environmental sustainability efforts. Increasing efficiency in the production process is not only economically beneficial but also helps to reduce the environmental impacts resulting from intensive farming practices, making it a more sustainable solution for the global palm oil industry.

***Author contributions statement*:** Yabani: Validation, Methodology, analysis of result., R.A.K: Writing, Investigation, conceptualization, A.S: Supervision project, validation., R.B.Y Syah: Orginal draft preparation, Data Curation.

***Conflict of interest*:** The author declares that there is no conflict of interests regarding the publication of this manuscript.

***Acknowledgment*:** Authors Thanks to Ministry Dikti DRTPM PDD Grant 020/LL1/AL.04.03/2024, 65/P3MPI/08.3.2/VI/2024 and PPKS Indonesian Oil Palm Research Institude (IOPRI).

***Funding*:** DRTPM PDD Grant 020/LL1/AL.04.03/2024, 65/P3MPI/08.3.2/VI/2024

***Ethical statement*:** All data used in this study was collected following standard industry practices and does not involve any ethical issues.

***Availability of data and material*:** The dataset is available upon request by email to the corresponding authors.

***Informed consent*:** Written consent was obtained from the participants to publish this data.

***Consent for publication*:** All authors submitted consent to publish this research article in JGIAS

***SDGs addressed*:** Zero Hunger, Responsible Consumption and Production, Climate Action.

REFERENCES

Ahmed, M., R. Seraj and S.M.S. Islam. 2020. The k-means algorithm: a comprehensive survey and performance evaluation. Electronics 9:1295.

Anselmi, A.A., J.P. Molin, H.C. Bazame and L. de P. Corrêdo. 2021. Definition of optimal maize seeding rates based on the potential yield of management zones. Agriculture 11:911.

Arevalo-Ramirez, T. and F. Auat Cheein. 2023. Cluster analysis for agriculture pp. 148-155. In Encyclopedia of digital agricultural technologies. Springer International Publishing.

Borlea, I.-D., R.-E. Precup and A.-B. Borlea. 2022. Improvement of K-means cluster quality by post-processing resulted clusters. Procedia Computer Science 199:63-70.

Cisneros, E., K. Kis-Katos and N. Nuryartono. 2021. Palm oil and the politics of deforestation in Indonesia. Journal of Environmental Economics and Management 108:102453.

Escallón-Barrios, M., D. Castillo-Gomez, J. Leal, C. Montenegro and A.L. Medaglia. 2022. Improving harvesting operations in an oil palm plantation. Annals of Operations Research 314:411-449.

Ezugwu, A.E., A.M. Ikotun, O.O. Oyelade, L. Abualigah, J.O. Agushaka, C.I. Eke and A.A. Akinyelu. 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence 110:104743.

Gintoron, C.S., M.A. Mohammed, S.N. Sazali, E.Q. Deka, K.H. Ong, I. H. Shamsi and P.J.H. King. 2023. Factors affecting pollination and pollinators in oil palm plantations: A review with an emphasis on the Elaeidobius kamerunicus weevil (Coleoptera: Curculionidae). Insects 14:454.

Govender, P. and V. Sivakumar. 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research 11:40-56.

Gupta, A. and P. Nahar. 2023. Classification and yield prediction in smart agriculture system using IoT. Journal of Ambient Intelligence and Humanized Computing 14:10235-10244.

Hassan, M.A., M.A.A. Farid, M.R. Zakaria, H. Ariffin, Y. Andou and Y. Shirai. 2024. Palm oil expansion in Malaysia and its countermeasures through policy window and biorefinery approach. Environmental Science & Policy 153:103671.

Ikotun, A.M., M.S. Almutari and A.E. Ezugwu. 2021. K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions. Applied Sciences 11:11246.

John Martin, J.J., R. Yarra, L. Wei and H. Cao. 2022. Oil palm breeding in the modern era: Challenges and opportunities. Plants 11:1395.

Joshua, V., S.M. Priyadharson and R. Kannadasan. 2021. Exploration of machine learning approaches for paddy yield prediction in eastern part of Tamilnadu. Agronomy 11:2068.

Li, K., T. Tscharntke, B. Saintes, D. Buchori and I. Grass. 2019. Critical factors limiting pollination success in oil palm: A systematic review. Agriculture, Ecosystems & Environment 280:152-160.

Li, M., Q. Fu, V.P. Singh, D. Liu, T. Li and Y. Zhou. 2020. Managing agricultural water and land resources with tradeoff between economic, environmental, and social considerations: A multi-objective non-linear optimization model under uncertainty. Agricultural Systems 178:102685.

Murphy, D.J., K. Goggin and R.R.M. Paterson. 2021. Oil palm in the 2020s and beyond: challenges and solutions. CABI Agriculture and Bioscience 2:39.

Niazian, M. and G. Niedbała. 2020. Machine learning for plant breeding and biotechnology. Agriculture 10:436.

Pathan, M., N. Patel, H. Yagnik and M. Shah. 2020. Artificial cognition for applications in smart agriculture: A comprehensive review. Artificial Intelligence in Agriculture 4:81-95.

Pathirana, R. and F. Carimi. 2022. Management and utilization of plant genetic resources for a sustainable agriculture. Plants 11:2038.

Qi, J., Y. Yu, L. Wang, J. Liu and Y. Wang. 2017. An effective and efficient hierarchical K-means clustering algorithm. International Journal of Distributed Sensor Networks 13:155014771772862.

Rousson, V. and N.F. Goşoniu. 2007. An -square coefficient based on final prediction error. Statistical Methodology 4:331-340.

Sarker, I.H. 2021. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science 2:160.

Seraj, A., M. Mohammadi-Khanaposhtani, R. Daneshfar, M. Naseri, M. Esmaeili, A. Baghban, S. Habibzadeh and S. Eslamian. 2023 pp. 89105. Cross-validation. In Handbook of Hydroinformatics Elsevier.

Sharma, A., A. Jain, P. Gupta and V. Chowdary. 2021. Machine learning applications for precision agriculture: A comprehensive review. IEEE Access 9:4843-4873.

Sharma, R., S.S. Kamble, A. Gunasekaran, V. Kumar and A. Kumar. 2020. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Computers & Operations Research 119:104926.

Sun, X., Q. Zhang, H. Zhang, L. Niu, M. Zhang and Y. Zhang. 2024. A set of artificial pollination technical measures: Improved seed yields and active ingredients of seeds in oil tree peonies. Plants 13:1194.

Sundaramoorthi, D. and L. Dong. 2022. Machine learning and optimization based decision-support tool for seed variety selection. Annals of Operations Research 341:5-39.

Swanson, B.E. and J.A. Huffman. 2020. Pollen clustering strategies using a newly developed single-particle fluorescence spectrometer. Aerosol Science and Technology 54:426-445.

Yabani, Y., R.A. Kuswardani, A. Susanto and R. Syah. 2023. Production approach model for oil palm (Elaeis guineensis Jacq) superior plant materials using artificial pollination studies. Journal of Global Innovations in Agricultural Sciences 11:587-594.

Yousefi, D.B., M., A.S. Mohd Rafie, S. Abd Aziz, S. Azrad, M. Mazmira Mohd Masri, A. Shahi and O.F. Marzuki. 2021. Classification of oil palm female inflorescences anthesis stages using machine learning approaches. Information Processing in Agriculture 8:537-549.

Yousefi, M., A.S. Mohd Rafie, S. Abd Aziz, S. Azrad and A. binti ABD Razak. 2020. Introduction of current pollination techniques and factors affecting pollination effectiveness by Elaeidobius kamerunicus in oil palm plantations on regional and global scale: A review. South African Journal of Botany 132:171-179.

Zhang, X. and C.-A. Liu. 2023. Model averaging prediction by fold cross-validation. Journal of Econometrics 235:280-301.