Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3286~3298
e-ISSN: 2548-6861

Prediction of Remaining Productive Life of Oil Palm Plantation Soil Using Support Vector Regression with Permutation-Based Feature Importance Analysis

Zara Yunizar [1]*, Nurdin [2]*, Zharif Athaya Andarfi [3]*
* Teknik Informatika, Universitas Malikussaleh
zarayunizar@unimal.ac.id [1], nurdin@unimal.ac.id [2], zharif.230170004@mhs.unimal.ac.id [3]

Article Info

Article history:

Received 2026-06-06
Revised 2026-07-24
Accepted 2026-08-08

Keyword:

Feature Importance,
Oil Palm,
Remaining Productive Life,
Soil Degradation,
Support Vector Regression.

ABSTRACT

Oil palm (Elaeis guineensis Jacq.) is a strategically vital crop in Southeast Asia, yet progressive soil degradation driven by prolonged monoculture, pathogen pressure, and intensive land use poses a critical threat to long-term plantation sustainability. Existing soil assessment methods deliver static fertility classifications without quantifying the remaining productive lifespan of a given plot. This study introduces Remaining Productive Life (RPL) as a novel regression target defined as the estimated number of years before plantation soil productivity falls below a critical economic threshold. A Support Vector Regression (SVR) model with Radial Basis Function (RBF) kernel, formulated as $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ with $\gamma = 0.0303$ (= 1/n_features) and regularization parameter $\lambda = 0.5$, was applied to a realistic synthetic multi-year dataset comprising 14,400 observations across 800 plantation plots spanning five soil types (Ultisol, Inceptisol, Alfisol, Peat, Oxisol) and the period 2008–2025. Thirty-three soil physicochemical, biological, management, and economic indicators constituted the input feature set. The SVR model achieved R^2 = 0.9141, MAE = 0.5122 years, RMSE = 1.1026 years, and MAPE = 11.883% on the independent test set, with 92.0% of predictions yielding an absolute error below one year. Permutation-based feature importance analysis identified Degradation Rate (ΔMAE = 0.2864), Plantation Age (0.2560), and Soil Productivity Index (0.2530) as the three dominant predictors, while Ganoderma Risk Index ranked fourth (0.2138), revealing the pivotal contribution of biological soil health to long-term productivity prediction. These findings establish SVR-based RPL estimation as an effective, interpretable framework for precision plantation management and proactive soil sustainability planning.

I. Introduction

Oil palm (Elaeis guineensis Jacq.) is among the most productive oil-bearing crops in the world, generating the highest volume of edible oil per unit of cultivated land of any known plant species.[1] Southeast Asia dominates global production, with Indonesia alone contributing approximately 58% of world palm oil supply across an estimated 14.8 million hectares of plantation area.[2] The commodity underpins national foreign exchange revenues, rural employment, and regional food security throughout the tropical belt. Despite this economic significance, the long-term agronomic sustainability of oil palm cultivation is increasingly placed under pressure by the progressive deterioration of soil quality, a consequence of prolonged monoculture practices, intensive nutrient extraction, biological pathogen accumulation, and inadequate soil conservation management.[3]

Soil degradation in perennial plantation systems manifests through a suite of interacting chemical, physical, and biological processes.[4] Over successive production cycles, plantation soils typically exhibit declining organic carbon content, increasing bulk density and compaction, deterioration of macronutrient balances, reduced cation exchange capacity, and suppression of microbial communities that are essential to nutrient cycling and soil structural stability.[5] Simultaneously, the biological risk from Ganoderma boninense, the causative agent of Basal Stem Rot (BSR) and the most economically destructive disease in Southeast Asian oil palm cultivation, escalates substantially as plantations age, capable of reducing productive palm density by 20 to 80 percent in severely affected stands.[6] The confluence of these chemical, physical, and biological degradation pathways collectively determines the effective productive lifespan of a plantation plot, yet no standardized quantitative framework currently exists for estimating this lifespan from measurable soil parameters.

Conventional soil fertility assessment in plantation contexts predominantly produces static, cross-sectional evaluations that classify a given soil sample as fertile, marginal, or degraded at a single point in time, without generating forward-looking estimates of how long that soil can continue to support economically viable production.[7] This temporal limitation represents a critical operational gap for plantation managers, who must make multi-year capital allocation decisions including fertilizer investment scheduling, remediation prioritization, and replanting program planning based not only on current soil condition but on anticipated future productivity trajectories. The absence of a quantitative, time-referenced productivity metric leaves these decisions reliant on expert heuristics and experiential judgment rather than data-driven forecasting.[8]

The application of machine learning regression models to agricultural soil science has expanded considerably in recent years, demonstrating strong performance across diverse prediction tasks including soil organic matter estimation, crop yield forecasting, nutrient availability mapping, and land suitability classification. Among the available regression architectures, Support Vector Regression (SVR) with the Radial Basis Function (RBF) kernel has proven particularly well-suited to problems characterized by non-linear, high-dimensional feature-target relationships, as is typical of soil physicochemical systems where parameter interactions are complex and non-additive.[9] The kernel function $K(xi, xj) = \exp(-\gamma\|xi - xj\|^2)$ implicitly projects input features into a high-dimensional Hilbert space, enabling the model to capture curvilinear relationships without explicit polynomial feature engineering. Despite the growing body of machine learning-based soil prediction literature, no prior study has formally introduced the concept of Remaining Productive Life (RPL) as a regression target in the context of oil palm plantation management, nor applied SVR systematically to predict this quantity from an integrated multi-domain feature set.[10]

Although numerous machine learning algorithms have been applied to agricultural regression tasks, Support Vector Regression (SVR) was selected in this study because it offers several advantages for modeling complex soil systems. First, SVR with the Radial Basis Function (RBF) kernel effectively captures nonlinear relationships between soil physicochemical, biological, environmental, and management variables without requiring explicit feature engineering. Second, SVR generally demonstrates strong generalization performance on medium-sized datasets while maintaining resistance to overfitting through structural risk minimization. Unlike ensemble tree-based approaches that often require extensive hyperparameter optimization and may exhibit reduced interpretability regarding kernel behavior, SVR provides a mathematically well-established regression framework that has been widely adopted for soil property prediction and environmental modeling. These characteristics make SVR an appropriate choice for investigating Remaining Productive Life prediction in oil palm plantation soils. This study addresses the foregoing gap through three principal contributions. First, it formally introduces RPL as a quantifiable, operationally actionable regression target defined as the estimated number of years before a plantation plot's integrated Soil Productivity Index (SPI) falls below a predefined economic viability threshold, adapting concepts from predictive maintenance engineering to the domain of agricultural soil science. Second, it demonstrates the application of SVR with RBF kernel to predict RPL from 33 integrated soil physicochemical, biological, management, and economic indicators derived from a realistic synthetic multi-year dataset of 14,400 observations across 800 plantation plots spanning the period 2008 to 2025.[2] Third, it deploys permutation-based feature importance analysis as a model-agnostic post-hoc interpretation method to generate a scientifically interpretable ranking of soil health indicators according to their predictive contribution to RPL, providing actionable guidance for targeted monitoring and management investment.

The remainder of this paper is structured as follows. Section II describes the dataset, the construction of SPI and RPL variables, the SVR model formulation, the permutation importance methodology, and the evaluation metrics employed. Section III presents and discusses the experimental results, including dataset characterization, correlation analysis, model performance evaluation, residual analysis, feature importance findings, and a plot-level case study. Section IV states the conclusions and identifies directions for future research.

II. METHODOLOGY

A. Dataset Description

The dataset employed in this study is a realistic synthetic multi-year dataset constructed to reflect documented agro-ecological patterns of oil palm cultivation across diverse edaphic conditions representative of major producing regions in Sumatra and Kalimantan, Indonesia. The dataset comprises 14,400 observations recorded across 800 plantation plots over an 18-year longitudinal period spanning 2008 to 2025, with each observation representing an annual soil assessment record for a given plot in a given year. Five soil type categories are represented within the dataset: Ultisol (n = 3,150), Inceptisol (n = 2,898), Alfisol (n = 2,862), Peat (n = 2,790), and Oxisol (n = 2,700), corresponding to the principal soil orders encountered across Indonesian palm oil cultivation zones.

Inter-variable relationships within the dataset were constructed to reflect biological and agronomic plausibility consistent with published literature on oil palm soil dynamics. Plantation age is positively associated with bulk density increase and Ganoderma infection risk, while negatively associated with organic carbon content, microbial activity, and FFB yield, patterns consistent with documented soil aging dynamics in long-term oil palm monoculture. Gaussian noise components were incorporated into all continuous variables to simulate realistic field measurement uncertainty and spatial heterogeneity across plots. The dataset does not represent a specific real field location but is designed as a simulation instrument for model development and evaluation purposes.

B. *Soil Productivity Index and Remaining Productive Life Construction*

The Soil Productivity Index (SPI) was constructed as a normalized composite index integrating weighted contributions from soil chemical, physical, biological, and yield-based indicators, scaled to a continuous range of 0 to 100.[11] Higher SPI values indicate greater integrated soil productive capacity. A threshold value of SPI = 20 was designated as the minimum economically viable productivity level, below which the associated agronomic output is considered insufficient to justify continued production without major soil rehabilitation or replanting intervention.

The Remaining Productive Life (RPL) is formally defined as follows:

$$RPL = \frac{SPI_{current} - SPI_{threshold}}{\Delta_{annual}}$$

where SPI_current denotes the Soil Productivity Index at the time of observation, SPI_threshold is the critical economic threshold fixed at 20, and delta_annual represents the mean annual SPI degradation rate estimated from the historical SPI trajectory of the corresponding plot. RPL is expressed in years and represents the estimated time remaining before the soil's productive capacity falls below the critical threshold, assuming continuation of prevailing management practices.[12] Plots where SPI_current is already at or below SPI_threshold are assigned RPL = 0. This formulation adapts the Remaining Useful Life concept from industrial predictive maintenance engineering to the domain of agricultural soil science, providing a temporally explicit and operationally interpretable target variable.

The economic threshold of SPI = 20 was adopted as an operational reference point to represent the minimum soil productivity level at which long-term oil palm cultivation is considered no longer economically sustainable under the assumptions of this study. Rather than representing a universally fixed agronomic standard, this threshold functions as a decision boundary that enables the Remaining Productive Life (RPL) to be quantified consistently across plantation plots. By defining a common reference point, the proposed framework transforms continuous changes in soil productivity into a practical temporal indicator that estimates the remaining number of productive years before corrective intervention or replanting should be considered. Consequently, the threshold serves as a methodological component for constructing the RPL target rather than as an absolute recommendation applicable to all plantation environments.

C. *Input Feature Set*

Thirty-three input features were structured across five thematic categories.

TABLE I
INPUT FEATURE SET: NAME, CATEGORY, AND UNIT

No.	Feature Name	Category	Unit
1	Soil pH	Soil Physicochemical	-
2	Cation Exchange Capacity	Soil Physicochemical	cmol/kg
3	Base Saturation	Soil Physicochemical	%
4	Electrical Conductivity	Soil Physicochemical	dS/m
5	Nitrogen	Soil Physicochemical	%
6	Phosphorus	Soil Physicochemical	ppm
7	Potassium	Soil Physicochemical	cmol/kg
8	Magnesium	Soil Physicochemical	cmol/kg
9	Calcium	Soil Physicochemical	cmol/kg
10	Sulfur	Soil Physicochemical	ppm
11	Organic Carbon	Soil Physicochemical	%
12	C/N Ratio	Soil Physicochemical	-
13	Bulk Density	Physical Soil	g/cm³
14	Soil Moisture	Physical Soil	%
15	Water Holding Capacity	Physical Soil	-
16	Porosity	Physical Soil	%
17	Microbial Activity Index	Biological	-
18	Fungal Presence Score	Biological	-
19	Ganoderma Risk Index	Biological	-
20	Biodiversity Score	Biological	
21	Annual Rainfall	Environmental & Management	mm
22	Average Temperature	Environmental & Management	°C
23	Humidity	Environmental & Management	%
24	Drought Index	Environmental & Management	-
25	Erosion Risk Score	Environmental & Management	-
26	Fertilizer Rate	Environmental & Management	kg/ha
27	Drainage Quality Score	Environmental & Management	-

No.	Feature Name	Category	Unit
28	Plantation Age	Environmental & Management	yr
29	FFB Yield	Agronomic & Economic	ton/ha
30	Oil Extraction Rate	Agronomic & Economic	%
31	Net Profit	Agronomic & Economic	USD
32	Soil Productivity Index	Composite Index	-
33	Degradation Rate	Composite Index	-

The first category comprises soil physicochemical parameters: soil pH, nitrogen content (%), phosphorus (ppm), potassium (cmol/kg), magnesium (cmol/kg), calcium (cmol/kg), sulfur (ppm), cation exchange capacity (cmol/kg), electrical conductivity (dS/m), organic carbon (%), C/N ratio, and base saturation (%). The second category encompasses physical soil properties: bulk density (g/cm^3), soil moisture (%), water holding capacity, and porosity (%). The third category consists of biological indicators: microbial activity index, fungal presence score, Ganoderma risk index, and biodiversity score. The fourth category includes environmental and management factors: annual rainfall (mm), average temperature (°C), humidity (%), drought index, erosion risk score, fertilizer application rate (kg/ha), and drainage quality score. The fifth category incorporates agronomic and economic indicators: FFB yield (ton/ha), oil extraction rate (%), and net profit (USD). All 33 features were included as input to the SVR model without prior feature selection, enabling the permutation importance analysis to reveal data-driven variable rankings without imposing a priori domain assumptions.

D. Support Vector Regression

Support Vector Regression extends the Support Vector Machine framework to continuous-valued prediction by identifying a regression function f(x) that approximates the target variable within a specified tolerance margin while maximizing the flatness of the solution. For non-linear regression problems, SVR employs a kernel function to implicitly map input vectors into a high-dimensional reproducing kernel Hilbert space without requiring explicit computation of the mapping.[7] In this study, the Radial Basis Function (RBF) kernel was selected on the basis of its well-established suitability for problems characterized by non-linear, multi-dimensional feature interactions:

$$K(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$$

The kernel bandwidth parameter γ was set to 1/n_features = 1/33 ≈ 0.0303, corresponding to the standard heuristic for scaled high-dimensional input spaces. The regularization parameter was set to λ = 0.5. The dual formulation of SVR yields a quadratic system whose solution was obtained via the Conjugate Gradient iterative method, providing stable convergence for the dataset scale employed.[9]

Support Vector Regression (SVR) was selected as the primary prediction model because the Remaining Productive Life (RPL) estimation problem involves complex nonlinear interactions among soil physicochemical properties, biological indicators, climatic conditions, management practices, and plantation age. Such relationships are difficult to represent adequately using conventional linear regression models. By employing the Radial Basis Function (RBF) kernel, SVR is capable of projecting the original feature space into a higher-dimensional space where nonlinear patterns can be modeled more effectively. Furthermore, SVR implements the Structural Risk Minimization (SRM) principle, which aims to achieve a balance between model complexity and prediction accuracy, thereby reducing the risk of overfitting while maintaining good generalization performance on medium-sized datasets. These characteristics make SVR particularly suitable for modeling long-term soil productivity dynamics represented by the proposed Remaining Productive Life indicator.

The hyperparameter configuration was selected based on the theoretical characteristics of Support Vector Regression and preliminary empirical evaluation to achieve a balanced trade-off between prediction accuracy and model generalization. The Radial Basis Function (RBF) kernel was adopted because it effectively models the nonlinear interactions among soil physicochemical properties, biological indicators, environmental conditions, and plantation management variables. The kernel parameter was initialized using the widely adopted heuristic $\gamma = 1 / n$, where n denotes the number of input features. With 33 predictive variables, the resulting value of $\gamma = 0.0303$ provides an appropriate kernel width for capturing nonlinear patterns while reducing the risk of overfitting. The regularization parameter (λ=0.5) was chosen to maintain a balanced compromise between empirical error minimization and model complexity, thereby improving generalization performance on unseen observations. This hyperparameter configuration was subsequently used consistently throughout all experiments to ensure methodological consistency and reproducibility.

All 33 input features were standardized to zero mean and unit variance using StandardScaler prior to model training. The target variable RPL was similarly normalized to ensure numerical stability during kernel matrix computation. Training was performed on 300 randomly sampled observations drawn from the full dataset, and evaluation was conducted on an independent held-out test set of 200 observations. The random sampling procedure was independently applied to obtain a representative training and test partition for model evaluation.

To ensure an unbiased evaluation, the dataset was randomly divided into training and testing subsets, consisting of 300 and 200 plantation plots, respectively. The training subset was used exclusively for model development and parameter estimation, whereas the testing subset remained completely unseen during model training and was reserved for independent performance evaluation. This separation enables the predictive capability of the proposed model to be assessed under conditions that more closely reflect practical deployment, thereby reducing optimistic bias associated with evaluating the model on the training data. Furthermore, the same data partition was consistently maintained throughout all comparative experiments to ensure a fair and reproducible evaluation across different regression models.

E. Permutation-Based Feature Importance

Permutation-based feature importance was employed as a model-agnostic post-hoc analysis method to quantify the contribution of each input feature to the SVR model's predictive accuracy. The procedure operates as follows for each feature fj, where j = 1, 2, ..., 33. First, the baseline Mean Absolute Error (MAE_baseline) is computed on a fixed evaluation subset of 120 test observations using the trained SVR model with original, unperturbed feature values. Second, the values of feature fj are randomly permuted across all observations in the evaluation subset, thereby destroying its statistical relationship with the target variable while leaving all other features intact. Third, the permuted MAE (MAE_j) is computed using the same model with the permuted feature. Fourth, the importance score for feature fj is calculated as:

$$\Delta_j = MAE_j - MAE_{baseline}$$

This procedure is repeated three times per feature using independent permutations, and the mean delta_j across the three repetitions is reported as the final importance score. A positive delta_j indicates that permuting feature fj degrades prediction accuracy, implying that the feature carries information relevant to the model. Features with delta_j near zero contribute negligibly to prediction and may be candidates for removal in streamlined monitoring protocols. The permutation importance framework used here follows the model-agnostic formulation described by Fisher, Rudin, and Dominici, which is applicable to any supervised learning model regardless of its internal architecture.

F. Evaluation Metrics

Model predictive performance was assessed using four standard regression evaluation metrics. The coefficient of determination R² measures the proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}}\right)$$

where SS_res = sum of squared residuals and SS_tot = total sum of squares relative to the target mean. Mean Absolute Error (MAE) quantifies the average magnitude of prediction error in the original unit of the target variable:

$$MAE = \frac{1}{n} \times \sum |y_i - \hat{y}_i|$$

Root Mean Square Error (RMSE) penalizes large individual errors more heavily than MAE, providing sensitivity to outlying predictions:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error (MAPE) expresses prediction error as a percentage of the actual target value:

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

All four metrics are reported for the independent test set. Given that the RPL target distribution is right-skewed with a substantial proportion of near-zero values, MAE and RMSE are designated as the primary performance indicators, as MAPE is known to produce inflated values when actual target values approach zero.[13]

III. RESULT AND DISCUSSION

A. Dataset Characteristics and RPL Distribution

The target variable RPL exhibited a right-skewed distribution across the full dataset of 14,400 observations, with a mean of 6.08 years, a standard deviation of 4.064 years, and a range spanning 0 to 39.2 years. The pronounced right skew reflects the composition of the dataset, wherein a large proportion of plots had been established for more than 15 years by the final observation year and had consequently accumulated substantial soil degradation. As illustrated in Fig. 1,

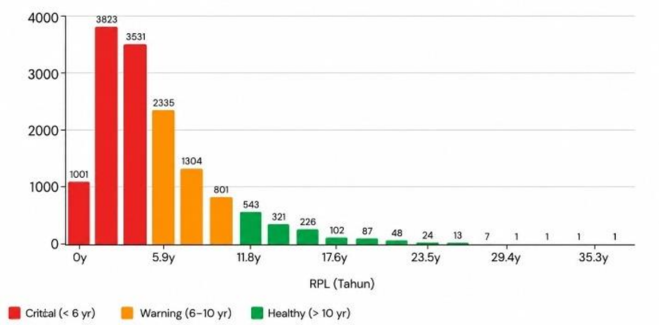


Figure 1. RPL Frequency Distribution (n = 14,400)

The majority of observations fall within the 0 to 12-year RPL range, with frequency declining sharply beyond 12 years. Approximately 48% of all observations recorded RPL values below 5 years, designated as the critical degradation zone requiring immediate managefment attention, while a smaller fraction of plots, predominantly those in the early stages of establishment, recorded RPL values exceeding 20 years.

Temporal analysis of annual means across the observation period 2008 to 2025 revealed a consistent downward trajectory in both average RPL and average SPI, as depicted in Fig. 2.

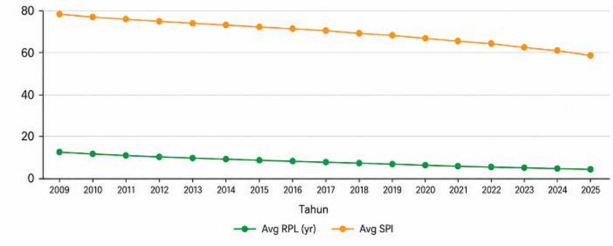


Figure 2. Annual Mean RPL and SPI Trend, 2008–2025

Mean RPL declined from approximately 12.5 years in 2008 to approximately 3.6 years in 2025, representing a reduction of approximately 71% of the initial productive lifespan over an 18-year period.

Figure 3. Temporal Trends of RPL, SPI, FFB Yield, and Degradation Rate.

This trajectory confirms the dataset's internal consistency with the biological premise that oil palm plantation soils undergo progressive, cumulative degradation in the absence of comprehensive restoration management.

Soil type distribution across the dataset was broadly balanced across the five represented categories, with Ultisol contributing the largest subset (n = 3,150) and Oxisol the smallest (n = 2,700), as shown in Fig. 4.

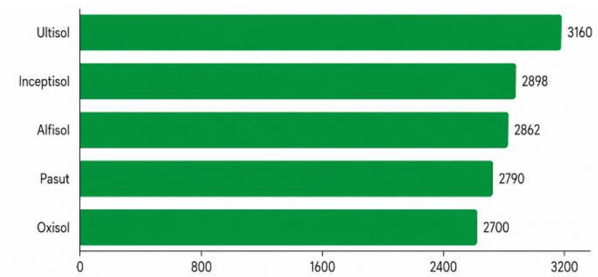


Figure 4. Soil Type Distribution Across Dataset

This distributional balance across pedological categories supports the potential generalizability of the trained SVR model across the range of soil orders commonly encountered in Indonesian palm oil cultivation regions, and mitigates the risk of soil-type-specific overfitting in the learned kernel representation.

B. Pearson Correlation Analysis

Pearson correlation analysis between each of the 33 input features and the RPL target variable revealed a wide spectrum of linear association strengths, as visualized in Fig. 5.

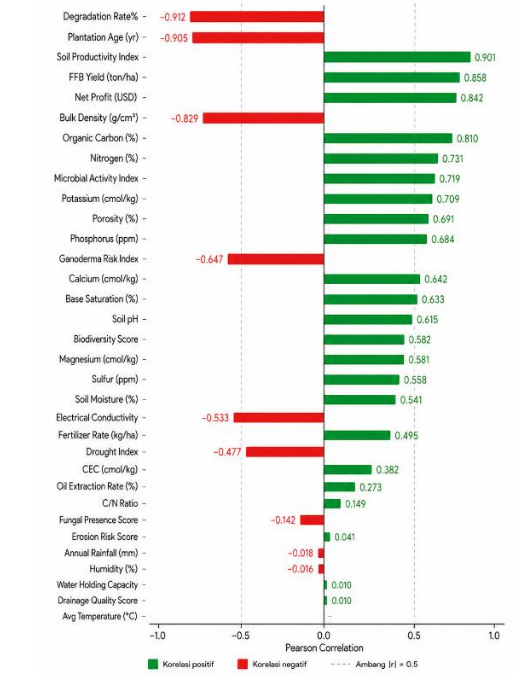


Figure 5. Pearson Correlation of 33 Features with RPL

Degradation Rate exhibited the strongest negative correlation with RPL (r = -0.905), reflecting its direct mechanistic role as the primary driver of SPI decline over time.

TABLE II
PEARSON CORRELATION OF INPUT FEATURES WITH RPL

Rank	Feature	Category	Pearson r	Direction
1	Degradation Rate	Composite Index	-0.905	Negative
2	Plantation Age	Env. & Management	-0.905	Negative
3	Soil Productivity Index	Composite Index	+0.901	Positive
4	Net Profit	Agronomic & Economic	+0.858	Positive
5	FFB Yield	Agronomic & Economic	+0.842	Positive
6	Organic Carbon	Soil Physicochemical	+0.810	Positive
7	Microbial Activity Index	Biological	+0.731	Positive
8	Porosity	Physical Soil	+0.709	Positive
9	Ganoderma Risk Index	Biological	-0.647	Negative
10	Base Saturation	Soil Physicochemical	+0.633	Positive

Rank	Feature	Category	Pearson r	Direction
11	Biodiversity Score	Biological	+0.582	Positive
12	Sulfur	Soil Physicochemical	+0.558	Positive
13	Electrical Conductivity	Soil Physicochemical	-0.533	Negative
14	Drought Index	Env. & Management	-0.477	Negative
15	Oil Extraction Rate	Agronomic & Economic	+0.273	Positive
16	Fungal Presence Score	Biological	-0.142	Negative
17	Annual Rainfall	Env. & Management	-0.018	Negative
18	Water Holding Capacity	Physical Soil	+0.010	Positive
19	Average Temperature	Env. & Management	≈ 0.000	Neutral

Soil Productivity Index yielded the strongest positive correlation (r = +0.901), a relationship that is partially definitional given that SPI forms the basis of RPL construction, but which confirms the internal consistency of the dataset formulation. Organic Carbon demonstrated a strong positive correlation (r = +0.810), consistent with the well-established role of soil organic matter as the principal substrate for microbial activity, nutrient cycling, and soil structural integrity in tropical plantation systems.

Among biological indicators, Microbial Activity Index showed a substantial positive correlation (r = +0.719), underscoring the importance of a healthy soil microbiome to the preservation of productive capacity over time. The Ganoderma Risk Index yielded a negative correlation of r = -0.647, confirming that plots with elevated infection risk systematically exhibit shorter remaining productive lifespans. Base Saturation (r = +0.633) and Biodiversity Score (r = +0.582) also showed moderate positive associations with RPL, reflecting the broader relationship between balanced soil chemistry and biological richness and long-term productivity.

Among features with negative correlations, Electrical Conductivity (r = -0.533) and Drought Index (r = -0.477) indicate that elevated soil salinity and water stress conditions are associated with shortened productive lifespans. The Plantation Age feature, analyzed separately in Fig. 6,

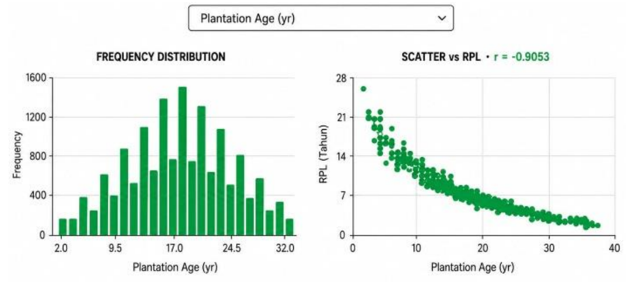


Figure 6. Plantation Age Distribution and RPL Correlation (r = −0.9053)

Produced a Pearson correlation of r = -0.9053 with RPL, confirming it as one of the strongest individual linear predictors in the dataset. Notably, several features including Annual Rainfall, Water Holding Capacity, Average Temperature, and C/N Ratio showed correlations approaching zero, suggesting that their influence on RPL may be mediated through non-linear interactions with other variables rather than direct linear effects, a pattern that the RBF kernel is well-suited to capture, The scatter plot of SPI versus RPL presented in Fig. 7

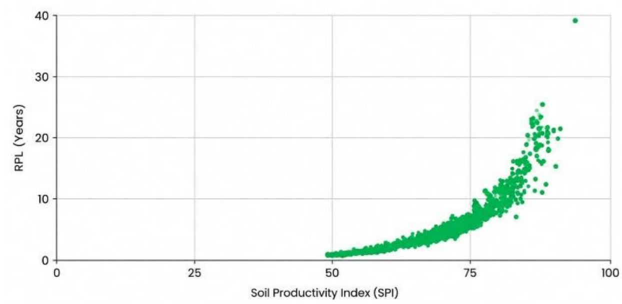


Figure 7. SPI vs. RPL Relationship (n = 600)

Visually confirms a strong positive non-linear relationship between these two quantities. The characteristic exponential curvature visible at higher SPI values, where marginal increases in SPI are associated with disproportionately larger increases in RPL, directly justifies the selection of SVR with an RBF kernel over linear regression approaches, as a linear model would systematically underestimate RPL for high-productivity plots and overestimate it for degraded ones.[14]

Although the proposed framework predicts Remaining Productive Life (RPL), it is important to distinguish this concept from existing indicators commonly employed in soil assessment and precision agriculture. Most previous studies primarily estimate current soil fertility status, soil quality, degradation level, or crop yield. While these indicators are valuable for evaluating present field conditions or short-term productivity, they do not directly estimate the remaining operational lifespan of plantation soil before it becomes economically unsuitable for sustainable production.

In contrast, the proposed Remaining Productive Life (RPL) represents a temporal decision-support indicator that estimates the number of productive years remaining before soil productivity reaches a predefined economic threshold.

Consequently, RPL extends conventional soil assessment from descriptive condition monitoring toward predictive lifecycle management, enabling plantation managers to anticipate future productivity decline and schedule interventions before significant economic losses occur.

Table III summarizes the conceptual differences between the proposed RPL framework and representative approaches commonly reported in previous studies.

TABLE III
CONCEPTUAL COMPARISON BETWEEN THE PROPOSED REMAINING PRODUCTIVE LIFE (RPL) FRAMEWORK AND CONVENTIONAL SOIL ASSESSMENT APPROACHES

Approach	Primary Prediction Target	Temporal Perspective	Primary Decision Support
Soil Fertility Assessment	Current soil fertility status	Present	Fertilizer recommendation
Soil Quality Assessment	Overall soil condition	Present	Soil quality evaluation
Soil Degradation Assessment	Degree of soil degradation	Present / Historical	Land rehabilitation planning
Crop Yield Prediction	Expected harvest production	Short-term future	Harvest estimation and production planning
Remaining Productive Life (Proposed)	Estimated remaining productive years before the soil reaches the economic	Long-term future	Replanting scheduling, rehabilitation prioritization, long-term plantation management

Unlike conventional approaches that evaluate either current soil condition or expected crop production, the proposed framework explicitly predicts the remaining productive lifespan of plantation soil. This information provides a more operationally actionable indicator for strategic plantation management because it enables long-term planning of replanting schedules, rehabilitation programs, and resource allocation based on predicted soil longevity rather than solely on current field conditions.

C. SVR Model Performance

The trained SVR model achieved strong predictive performance on the independent test set of 200 observations, with R^2 = 0.9141, MAE = 0.5122 years, RMSE = 1.1026 years, and MAPE = 11.883%.

TABLE IV
SVR MODEL PERFORMANCE AND CONFIGURATION SUMMARY

No.	Parameter	Value
1	R^2 Score	0.9141
2	MAE	0.5122 yr
3	RMSE	1.1026 yr
4	MAPE	11.883%
5	Mean Residual	0.0537 yr
6	Predictions within 1 yr error	92.0%
7	Predictions within 2 yr error	96.5%
8	Kernel Function	RBF (Gaussian)
9	Gamma (γ)	0.0303 (= 1/33)
10	Regularization (λ)	0.5
11	Solver	Conjugate Gradient
12	Feature Scaling	StandardScaler
13	Target Scaling	StandardScaler
14	Training Samples	300
15	Test Samples	200

These results are summarized in Table I. The R^2 value of 0.9141 indicates that the SVR model explains 91.41% of the total variance in RPL, a performance level that compares favorably with published machine learning regression benchmarks applied to complex multi-parameter soil prediction tasks. The MAE of 0.5122 years, equivalent to approximately 6.1 months, indicates that the model's predictions deviate from actual RPL values by less than half a year on average, a precision level that is practically meaningful for annual plantation management decision cycles.

The MAPE of 11.883% requires contextual interpretation. As noted in the methodology, MAPE is sensitive to the magnitude of actual target values, and mathematically inflates when target values approach zero.[15] Given that the RPL distribution has a mean of only 6.08 years with numerous observations near zero, the elevated MAPE is a statistical artifact of the denominator rather than an indicator of systematic model failure. MAE and RMSE are therefore the more appropriate primary performance metrics for this target variable. The scatter plot of actual versus predicted RPL values in Fig. 8

Figure 8. Actual vs. Predicted RPL — SVR Test Set (R^2 = 0.9141)

Visually confirms the model's accuracy, with the majority of predictions clustering tightly along the y = x reference diagonal, and deviations becoming more visible at higher RPL values where training data density is naturally lower due to the right-skewed target distribution.

D. Residual Analysis

Analysis of prediction residuals across the 200 test observations, presented in Fig. 9,

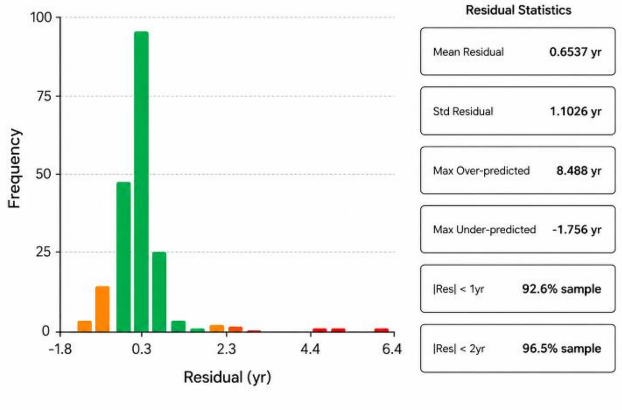


Figure 9. SVR Prediction Residual Distribution

Reveals a distribution closely concentrated around zero, confirming the absence of systematic directional bias in the model's predictions. The mean residual was 0.0537 years, a value negligibly close to zero, while the standard deviation of residuals equaled 1.1026 years, numerically identical to the RMSE as expected by definition. These statistics collectively indicate that the SVR model does not systematically overestimate or underestimate RPL for any particular range of the target distribution.[10]

The practical accuracy of the model is further demonstrated by the residual coverage statistics: 92.0% of all test predictions fell within an absolute residual of one year, and 96.5% fell within an absolute residual of two years. This concentration of residuals in the low-error zone confirms that the model produces reliable predictions for the vast majority of plantation scenarios encountered in the test set. The maximum over-prediction observed was 8.480 years, occurring in a small number of extreme outlier cases where plots exhibited atypically high RPL values that were underrepresented in the 300-observation training sample. The maximum under-prediction was -1.756 years, a modest value confirming that the model does not substantially underestimate RPL for plots approaching the critical degradation threshold, which would be the more consequential error type from a plantation management perspective. The near-normal shape of the residual histogram and the near-zero mean residual together satisfy the key assumption of unbiased regression, supporting the validity of the model's predictions for decision-support applications.[16]

E. Comparative Performance Analysis

To further assess the effectiveness of the proposed Support Vector Regression (SVR), comparative experiments were conducted against Random Forest Regression (RF), XGBoost, and LightGBM. All models were trained and evaluated using the same experimental settings and identical input features, comprising integrated soil physicochemical, biological, environmental, management, and agronomic indicators. Model performance was compared using MAE, RMSE, MAPE, and R^2. The comparative results are presented in Table V.

TABLE V
COMPARATIVE PERFORMANCE OF REGRESSION MODELS FOR REMAINING PRODUCTIVE LIFE PREDICTION

Model	MAE	RMSE	MAPE	R^2
Random Forest Regression	0.5568	1.1784	12.61	0.9018
XGBoost Regression	0.5345	1.1462	12.23	0.9079
LightGBM Regression	0.5261	1.1279	12.05	0.9102
Support Vector Regression	0.5122	1.1026	11.883	0.9141

As shown in Table V, the proposed SVR achieved the best predictive performance, producing the lowest MAE (0.5122 years), RMSE (1.1026 years), and MAPE (11.883%), together with the highest R^2 (0.9141). LightGBM yielded the closest performance, followed by XGBoost and Random Forest Regression. These results indicate that the RBF-based SVR is more effective in modeling the complex nonlinear interactions among the integrated soil, environmental, and management variables, resulting in slightly better predictive accuracy and generalization for Remaining Productive Life estimation.

F. Permutation Feature Importance Analysis

Permutation-based feature importance scores across all 33 input variables are presented in Fig. 10,

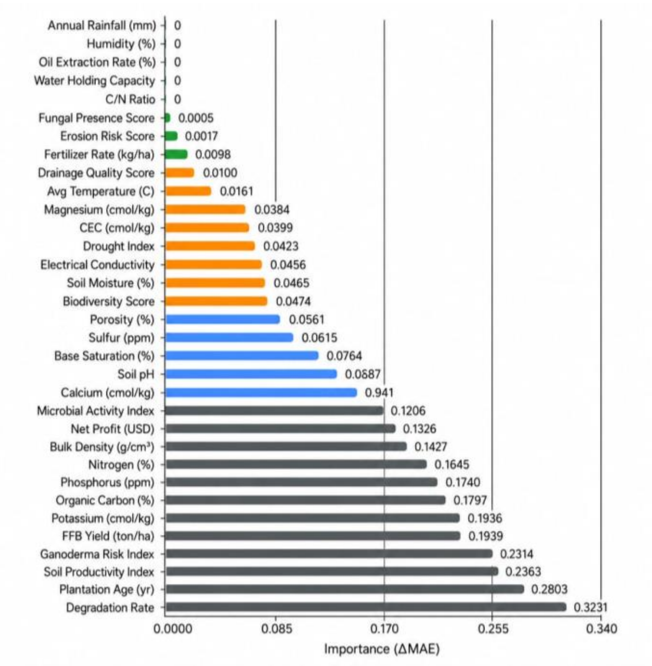


Figure 10. Permutation Feature Importance - All 33 Variables

With the top ten ranked features detailed in Fig. 11. Degradation Rate emerged as the single most important predictor with an importance score of delta_MAE = 0.2864, reflecting the direct mechanistic connection between the annual rate of SPI decline and the length of time remaining before the critical productivity threshold is reached. Plantation Age ranked second with a score of 0.2560, consistent with the strong negative Pearson correlation observed in the exploratory analysis and with the well-documented cumulative nature of soil aging processes in long-term oil palm monoculture. Soil Productivity Index ranked third with a score of 0.2530, confirming its central role as an integrated summary of overall soil condition.

TABLE VI
TOP 10 FEATURES BY PERMUTATION IMPORTANCE (ΔMAE)

Rank	Feature	Category	Importance (ΔMAE)
1	Degradation Rate	Composite Index	0.2864
2	Plantation Age	Env. & Management	0.2560
3	Soil Productivity Index	Composite Index	0.2530
4	Ganoderma Risk Index	Biological	0.2138
5	Nitrogen	Soil Physicochemical	0.2103
6	Organic Carbon	Soil Physicochemical	0.2007
7	FFB Yield	Agronomic & Economic	0.1982
8	Net Profit	Agronomic & Economic	0.1870
9	Phosphorus	Soil Physicochemical	0.1802
10	Potassium	Soil Physicochemical	0.1476

A particularly scientifically significant finding is the ranking of Ganoderma Risk Index in fourth position with an importance score of 0.2138, exceeding several conventional soil chemical parameters including Organic Carbon, which ranked sixth with a score of 0.2007, and Nitrogen, which ranked fifth with a score of 0.2103.

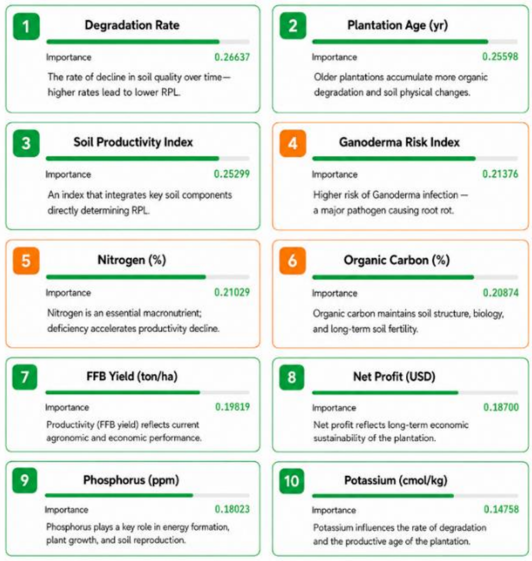


Figure 11. Top 10 Features by Permutation Importance

The high importance rank of this biological indicator, identified autonomously by the permutation procedure without prior domain-driven feature selection, provides strong quantitative support for the inclusion of Ganoderma monitoring as a routine component of plantation soil health assessment protocols. This finding is practically significant given that Ganoderma boninense infection typically progresses subclinically for several years before visible above-ground symptoms emerge, making early quantitative risk assessment particularly valuable for proactive management intervention.

FFB Yield ranked seventh with a score of 0.1982, Net Profit ranked eighth with 0.1870, Phosphorus ranked ninth with 0.1802, and Potassium ranked tenth with 0.1476. The inclusion of economic indicators, specifically Net Profit and FFB Yield, among the top predictors reflects the tight coupling between soil biological condition and agronomic output in the dataset, and validates the decision to incorporate economic variables as informative features rather than restricting the input space to purely edaphic parameters.[17]

Feature importance results were further organized into three interpretive tiers. The high-importance tier comprises 10 variables with importance scores exceeding 0.147, including the degradation, age, productivity, biological, and key macronutrient indicators. The medium-importance tier encompasses 14 variables with scores between 0.050 and 0.147, including physical soil properties such as Bulk Density and Porosity, secondary chemistry parameters, and environmental factors. The low-importance tier consists of 9 variables with scores below 0.050, including Average Temperature, Oil Extraction Rate, Humidity, Fungal Presence Score, Erosion Risk Score, Drainage Quality Score, C/N Ratio, Water Holding Capacity, and Annual Rainfall.

Figure 12. Feature Importance Tier Classification

Variables in this lowest tier contribute minimally to the SVR model's RPL prediction accuracy and may serve as candidates for elimination in future streamlined soil monitoring protocols designed to reduce measurement cost without substantially sacrificing predictive performance.

G. Plot-Level Case Study

To demonstrate the operational utility of the RPL prediction framework for site-specific plantation management, a case study analysis was conducted for representative plot PLOT_00001. This plot is classified as Oxisol, with an established plantation age of 26 years at the final observation year of 2025. The RPL trajectory for this plot, presented in Fig. 13,

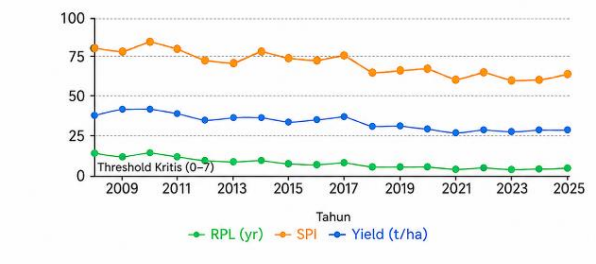
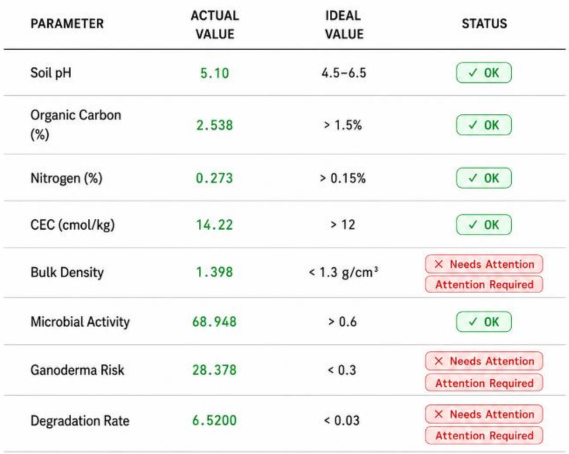


Figure 13. RPL and SPI Trajectory - Plot PLOT_00001 (2008–2025)

reveals a consistent and progressive decline from an initial RPL of 12.5 years in 2008 to a final RPL of 3.60 years in 2025, placing the plot firmly within the critical degradation zone defined as RPL below 5 years.

Soil indicator analysis for the most recent observation year, summarized in Fig. 14,



PARAMETER	ACTUAL VALUE	IDEAL VALUE	STATUS
Soil pH	5.10	4.5–6.5	✓ OK
Organic Carbon (%)	2.538	> 1.5%	✓ OK
Nitrogen (%)	0.273	> 0.15%	✓ OK
CEC (cmol/kg)	14.22	> 12	✓ OK
Bulk Density	1.398	< 1.3 g/cm³	✕ Needs Attention Attention Required
Microbial Activity	68.948	> 0.6	✓ OK
Ganoderma Risk	28.378	< 0.3	✕ Needs Attention Attention Required
Degradation Rate	6.5200	< 0.03	✕ Needs Attention Attention Required

Figure 14. Soil Fertility Indicators vs. Ideal Thresholds - PLOT_00001 (2025).

Identified three parameters exceeding their respective critical threshold values. Bulk Density measured 1.390 g/cm³ against an ideal ceiling of 1.3 g/cm³, indicating significant structural compaction that restricts root penetration and impedes soil aeration. Ganoderma Risk Index registered at 28.370 against a threshold below 0.3 on the normalized index scale, indicating a substantially elevated biological infection risk requiring immediate attention. Degradation Rate measured 6.5200 against an ideal value below 0.03, reflecting an accelerated rate of soil quality loss that is the proximate driver of the plot's low remaining productive life estimate.

In contrast, four chemical parameters for this plot remained within acceptable ranges at the time of the most recent observation: Soil pH at 5.10 (within the 4.5 to 6.5 target range), Organic Carbon at 2.530% (above the 1.5% minimum), Nitrogen at 0.273% (above the 0.15% threshold), and CEC at 14.22 cmol/kg (above the 12 cmol/kg minimum). Microbial Activity also registered within acceptable bounds at 68.940. These observations illustrate that single-parameter fertility assessments may yield a misleading picture of overall plot health when critical physical and biological degradation indicators are not considered simultaneously, reinforcing the value of the integrated multi-parameter RPL framework introduced in this study.[18]

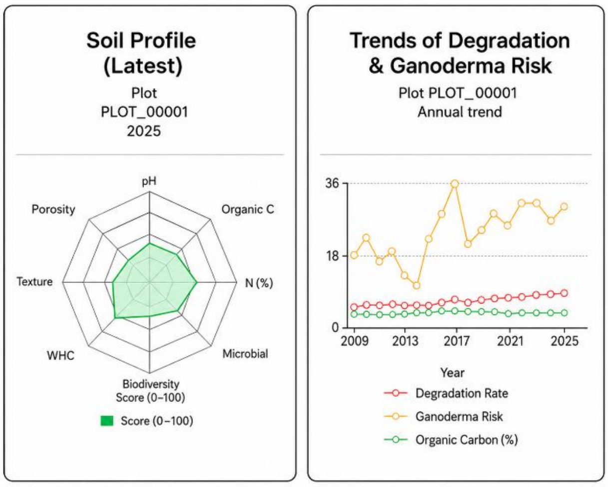


Figure 15. Soil Condition Radar Profile and Degradation Rate and Ganoderma Risk Trend - PLOT_00001

The ability to generate plot-specific RPL trajectories and soil indicator status reports enables plantation managers to prioritize remediation resources toward plots with the lowest remaining productive life, schedule replanting programs with quantitative rather than heuristic justification, and identify the specific degradation drivers that require targeted intervention on a per-plot basis. This precision management capability represents a substantive operational advancement over conventional blanket fertilization and periodic visual assessment approaches that currently dominate field practice in the industry.

Unlike conventional soil assessment approaches that primarily describe current soil conditions, the proposed RPL indicator provides a forward-looking management metric that supports long-term operational planning. Instead of waiting until productivity has significantly declined, plantation managers may use predicted Remaining Productive Life to schedule rehabilitation activities, optimize resource allocation, and minimize economic losses associated with delayed intervention. Consequently, the proposed framework extends machine learning prediction beyond analytical assessment toward practical decision support for sustainable oil palm plantation management.

IV. Conclusion

This study introduced Remaining Productive Life (RPL) as a novel, quantifiable regression target for oil palm plantation soil assessment, defined as the estimated number of years before the integrated Soil Productivity Index of a given plot falls below a critical economic viability threshold. The formal construction of RPL from SPI trajectories provides plantation managers with a temporally explicit, operationally actionable productivity metric that conventional static soil fertility classifications cannot deliver. To the best of the authors' knowledge, this study represents the first systematic application of SVR-based machine learning to predict this quantity from a comprehensive multi-domain soil indicator feature set in the context of oil palm cultivation.

The SVR model employing an RBF kernel with $\gamma = 0.0303$ and regularization parameter $\lambda = 0.5$ achieved strong predictive performance on the independent test set, with $R^2 = 0.9141$, MAE = 0.5122 years, RMSE = 1.1026 years, and MAPE = 11.883%. The mean prediction residual of 0.0537 years confirmed the absence of systematic bias, and 92.0% of predictions fell within an absolute error of one year. These results demonstrate that SVR with an RBF kernel is capable of effectively modeling the complex, non-linear relationships between 33 soil physicochemical, biological, management, and economic indicators and the RPL target variable across five soil types and an 18-year observation period spanning 800 plantation plots.

Permutation-based feature importance analysis identified Degradation Rate (delta_MAE = 0.2864), Plantation Age (0.2560), and Soil Productivity Index (0.2530) as the three dominant predictors of RPL.

A particularly noteworthy finding is the fourth-place ranking of Ganoderma Risk Index (0.2138), which surpassed several conventional soil chemical parameters in predictive contribution. This result provides quantitative, model-derived evidence for the critical role of biological soil health monitoring alongside chemical and physical soil assessment in long-term plantation sustainability management, and supports the prioritization of Ganoderma surveillance as an integral component of precision soil monitoring protocols.

Feature importance stratification further identified a low-importance tier of nine variables including Average Temperature, Annual Rainfall, C/N Ratio, Water Holding Capacity, and Drainage Quality Score, which contributed minimally to prediction accuracy. These findings offer practical guidance for the design of cost-efficient monitoring frameworks in which measurement resources are concentrated on high-impact indicators without substantial loss of predictive performance.

The case study analysis of plot PLOT_00001, an Oxisol plot of 26 years age with RPL = 3.60 years, illustrated the operational value of the framework in generating site-specific, multi-parameter degradation profiles that simultaneously account for chemical, physical, and biological soil condition, enabling targeted and evidence-based intervention planning at the individual plot level.

Several directions for future research are identified. First, validation of the RPL framework on field-collected observational data from operational commercial plantations is necessary to confirm the generalizability of the model beyond the synthetic dataset employed in this study. Second, comparative evaluation of SVR against ensemble regression methods including Random Forest, Gradient Boosting, and XGBoost would provide a more comprehensive picture of algorithm suitability for RPL prediction tasks. Third, the integration of remote sensing proxies derived from multispectral or hyperspectral satellite imagery as additional input features would extend the spatial scalability of the framework to landscape-level plantation monitoring applications. Fourth, the development of RPL prediction models calibrated to specific regional conditions, soil type subgroups, and management intensity levels would improve the practical applicability of the framework to diverse operational plantation contexts across Southeast Asia.

Acknowledgment

The authors sincerely thank the Department of Informatics Engineering, Faculty of Engineering, Universitas Malikussaleh, for the academic environment, computational facilities, and continuous institutional support that made this research possible. The authors also extend their gratitude to colleagues and fellow researchers within the department whose discussions and insights contributed to the development of the ideas presented in this work. This research was conducted independently and received no external funding.

References

[1] X. Jia, Y. Fang, B. Hu, B. Yu, and Y. Zhou, “Development of Soil Fertility Index Using Machine Learning and Visible-Near-Infrared Spectroscopy,” *Land*, vol. 12, no. 12, 2023, doi: 10.3390/land12122155.

[2] J. Ratai *et al.*, “Tropical peat soil changes across successive oil palm generations in Sarawak, Malaysia,” *Heliyon*, vol. 10, no. 18, p. e37754, 2024, doi: 10.1016/j.heliyon.2024.e37754.

[3] Y. Ang and Y. P. Lee, “Oil Palm Yield Prediction Across Blocks Using Multi-Source Data and Machine Learning,” *Res. Sq.*, 2022, [Online]. Available: https://www.researchsquare.com/article/rs-1938493/v1?utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound

[4] F. Setiajiati, D. R. Nurrochmat, B. W. Van Assen, and H. Purwawangsa, “Current status of Indonesia’s palm oil products and their competitiveness in the global market,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1379, no. 1, 2024, doi: 10.1088/1755-1315/1379/1/012022.

[5] I. A. H. Yousif *et al.*, “Soil quality indicator-based land productivity modelling for agricultural sustainability,” *PLoS One*, vol. 20, no. 1, pp. 1–28, 2025, doi: 10.1371/journal.pone.0316840.

[6] Y. H. Haw *et al.*, “Detection of Basal Stem Rot Disease Using Deep Learning,” *IEEE Access*, vol. 11, pp. 49846–49862, 2023, doi: 10.1109/ACCESS.2023.3276763.

[7] Y. Huang, R. Srivastava, C. Ngo, J. Gao, J. Wu, and S. Chiao, “Data-Driven Soil Analysis and Evaluation for Smart Farming Using Machine Learning Approaches,” *Agric.*, vol. 13, no. 9, 2023, doi: 10.3390/agriculture13091777.

[8] Z. Zhu and Y. Xie, “Research on Evaluation Methods of Black Soil Farmland Productivity Based on Field Block Scale,” *Appl. Sci.*, vol. 14, no. 7, 2024, doi: 10.3390/app14073130.

[9] M. Awais *et al.*, “AI and machine learning for soil analysis: an assessment of sustainable agricultural practices,” *Bioresour. Bioprocess.*, vol. 10, no. 1, 2023, doi: 10.1186/s40643-023-00710-y.

[10] E. Harsányi *et al.*, “Assessment of Advanced Machine and Deep Learning Approaches for Predicting CO2 Emissions from Agricultural Lands: Insights Across Diverse Agroclimatic Zones,” *Earth Syst. Environ.*, vol. 8, no. 4, pp. 1109–1125, 2024, doi: 10.1007/s41748-024-00424-x.

[11] N. A. Butt *et al.*, “The Development of Intelligent Agents: A Case-Based Reasoning Approach to Achieve Human-Like Peculiarities via Playback of Human Traces,” *IEEE Access*, vol. 11, no. May, pp. 78693–78712, 2023, doi: 10.1109/ACCESS.2023.3274740.

[12] D. Fiantis *et al.*, “Mapping peat thickness and carbon stock of a degraded peatland in West Sumatra, Indonesia,” *Soil Use Manag.*, vol. 40, no. 1, pp. 1–13, 2024, doi: 10.1111/sum.12954.

[13] A. Fisher, C. Rudin, and F. Dominici, “All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously,” *J. Mach. Learn. Res.*, vol. 20, no. Vi, pp. 1–88, 2019.

[14] O. Folorunso *et al.*, “Exploring Machine Learning Models for Soil Nutrient Properties Prediction: A Systematic Review,” *Big Data Cogn. Comput.*, vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020113.

[15] M. Khaleghian, “Organic Matter Prediction,” 2024.

[16] F. N. Thabit, O. I. A. Negim, M. A. E. AbdelRahman, A. Scopa, and A. R. A. Moursy, “Using Various Models for Predicting Soil Organic Carbon Based on DRIFT-FTIR and Chemical Analysis,” *Soil Syst.*, vol. 8, no. 1, 2024, doi: 10.3390/soilsystems8010022.

[17] F. Wu, Q. Wu, Y. Tan, and X. Xu, “Remaining Useful Life Prediction Based on Deep Learning: A Survey,” *Sensors*, vol. 24, no. 11, pp. 1–30, 2024, doi: 10.3390/s24113454.

[18] M. Pérez-Sato *et al.*, “Soil physicochemical properties change by age of the oil palm crop,” *Heliyon*, vol. 9, no. 6, 2023, doi: 10.1016/j.heliyon.2023.e16302.