

Ridge Regression with Walk-Forward Validation and Explainable AI for Tea Yield Prediction in Sri Lanka

Kavishka Karunagaran
0009-0006-6800-0121

Dept. of Artificial Intelligence and Data Science
Informatics Institute of Technology
Colombo, Sri Lanka
Email: kavishka.20240218@iit.ac.lk

Abstract—Sri Lanka’s tea industry has faced significant yield declines due to climate variability and a lack of data-driven forecasting tools. This paper presents a machine learning-based yield prediction system using Ridge Regression with lag feature engineering and walk-forward validation on longitudinal data from a Sri Lankan regional estate (2017–2025). The system integrates fertilizer records with historical weather data from the Open-Meteo Archive API to forecast monthly Yield Per Hectare (YPH). Comparative experiments empirically established an optimal 3-year rolling training window. Temporally ordered walk-forward validation is utilized to eliminate data leakage and realistically simulate agricultural forecasting conditions. Model explainability is achieved intrinsically via direct Ridge coefficient analysis. The proposed model achieves an R^2 of 0.9201, outperforming CatBoost, XGBoost, Random Forest, and Neural Networks. While validated on a single estate, results confirm its robustness for deployment within Sri Lankan tea estate management systems.

Index Terms—tea yield prediction, Ridge Regression, walk-forward validation, explainable AI, lag feature engineering, precision agriculture, Sri Lanka, time-series forecasting, multi-source data fusion

I. INTRODUCTION

SRI Lanka’s tea industry, globally recognized under the Ceylon Tea brand, is a vital contributor to national foreign exchange earnings, rural employment, and economic stability. In 2024, domestic tea production declined by 14.95% compared to the previous year, the greatest crop shortfall recorded since 1995 with climate change and extreme weather conditions identified as primary drivers [1]. The Central Bank of Sri Lanka further reported a year-on-year production decrease of 7.5% as of October 2024 [2], underscoring the fragility of tea cultivation under shifting environmental conditions.

Despite the economic importance of the sector, tea yield forecasting in Sri Lanka has remained largely dependent on manual estimation and subjective expert judgment. A survey by Weththasinghe and Dissanayake (2024) revealed that 65.5% of Sri Lankan tea farmers possess poor ICT skills, yet 73.4% expressed willingness to adopt technology-based solutions [3].

This work was supported by the Informatics Institute of Technology

This represents a significant opportunity for AI-driven agricultural decision support systems that are both accurate and interpretable.

Existing machine learning research on tea yield prediction, while growing, suffers from three critical methodological gaps: (1) reliance on random train-test splits that ignore the temporal ordering of agricultural data and introduce data leakage; (2) omission of lag features that capture seasonal carryover effects and delayed agronomic responses; and (3) insufficient explainability, with many studies reporting only aggregate accuracy metrics rather than feature-level insights that farmers and estate managers can act upon.

This paper addresses these gaps by presenting a yield prediction system built on Ridge Regression trained with a 3-year rolling walk-forward validation strategy, enriched with lag features capturing temporal dependencies in yield and fertilizer usage. The model is trained on eight years of real estate data augmented with weather data. Explainability is achieved through direct Ridge coefficient analysis, which quantifies each feature’s proportional contribution to the predicted yield without requiring external post-hoc explanation libraries.

The key contributions of this work are: (i) the first Sri Lankan tea yield prediction model with explicit temporal modeling through walk-forward validation; (ii) a multi-source feature fusion approach combining fertilizer, weather, and lag variables; (iii) a model-intrinsic XAI approach using Ridge coefficients for direct feature attribution; and (iv) empirically validated window size selection demonstrating that 3 years optimally balances model stability and generalization for this domain.

II. RELATED WORK

Research in AI-driven crop yield prediction has expanded rapidly, though work specifically targeting Sri Lankan tea yield remains sparse. The most directly relevant study is by Mampitiya et al. (2025) [4], who developed an explainable AI framework for Sri Lankan tea yield prediction using CatBoost with SHAP-based feature attribution, achieving an R^2 of approximately 0.9037. While this represents important

progress in applying XAI to Sri Lankan tea agriculture, the dataset was partitioned using a random train-test split rather than a temporally ordered strategy. This approach does not account for temporal autocorrelation in agricultural yield data and may yield optimistically biased performance estimates. Furthermore, their model did not incorporate lag features to capture delayed agronomic effects such as the influence of prior-month fertilization or lagged seasonal yield patterns.

Deka et al. (2024) [5] compared statistical and machine learning models for tea yield prediction in Assam, India, using 23 years of climatic data and finding Random Forest to be the best performer. However, this study treated the prediction task as generic regression, applying standard cross-validation without temporal ordering, and excluded management variables such as fertilizer inputs. Batool et al. (2022) [6] proposed a hybrid framework combining FAO AquaCrop simulation with machine learning regressors with XGBoost achieving the best performance, at a Pakistani research station. Despite the innovative multi-model approach, the dataset was limited to a single site with a short time horizon, and no explicit temporal modeling or explainability analysis was provided.

Phan et al. (2022) [7] used multi-temporal MODIS NDVI satellite data alongside meteorological variables to predict seasonal tea yield in Vietnam, achieving R^2 of approximately 0.73 with Random Forest. While the remote sensing approach is scalable, it relies on coarse-resolution data and omits estate-level management practices, limiting local applicability. Jui et al. (2022) [8] extended this line with a spatiotemporal hybrid Random Forest model using satellite-derived variables, though similarly lacking temporal validation strategies.

To contextualize these within broader computational agriculture, related frameworks in the literature have explored specialized sub-problems, such as the automated construction of knowledge graphs for tracking tea diseases and pests [9], as well as the optimization of multi-level ensemble structures to stabilize macro-level crop yield forecasts [10].

Across these studies, a consistent pattern emerges: predictive performance is heavily influenced by dataset characteristics, feature selection, and the appropriateness of the validation strategy to the data structure. None of the reviewed studies combine temporal walk-forward validation, lag feature engineering, multi-source data fusion, and model-intrinsic explainability within a single system applied to Sri Lankan estate-level data, the gap this work addresses.

III. METHODOLOGY

A. Data Collection and Sources

The primary dataset was manually collected from a tea estate in Sri Lanka, covering the period from 2017 to 2025. Monthly records include tea yield measured in kilograms per hectare (YPH) and six fertilizer application variables: T200, U709, Epsom Salt, Dolomite, Urea, and BioShield. The estate identity is anonymized in accordance with data access agreements.

Daily weather data for the same geographic location and time period were retrieved from the Open-Meteo Archive

API [11], providing: mean daily temperature ($^{\circ}\text{C}$), total rainfall (mm), mean wind speed at 10 m (km/h), and reference evapotranspiration computed by the FAO-56 Penman-Monteith method (ET_0 , mm/day). Daily weather variables were aggregated to monthly resolution using monthly sums for rainfall and ET_0 , and monthly means for temperature and wind speed. These were then merged with the estate fertilizer-yield records to form the final modelling dataset of 96 monthly observations.

B. Feature Engineering

Three categories of features were constructed. First, raw agronomic and weather variables were retained as direct inputs. Second, a derived interaction feature, Dolomite Moisture (DM), was computed as:

$$\text{DM} = \text{Dolomite} \times (\text{rain_sum} - \text{ET}_0) \quad (1)$$

capturing the interaction between calcium-magnesium amendment application and effective soil moisture availability under evapotranspiration demand. Third, three lag features were engineered to model temporal dependencies inherent in tea yield: `YPH_lag1` (yield from the previous month), `YPH_lag12` (yield from the same month in the previous year, capturing seasonal periodicity), and `Compost_lag1` (compost application from the previous month, capturing delayed nutrient uptake).

The lag features are particularly important because agricultural yield exhibits strong autocorrelation: the harvest of one month is influenced by root health and soil conditioning from prior months. By explicitly encoding these temporal dependencies, the model captures dynamics that standard regression without lag features cannot represent. The complete feature set comprises 14 features: T200, U709, Epsom Salt, Dolomite, Urea, BioShield, `temperature_2m_mean`, `rain_sum`, `wind_speed_10m_mean`, ET_0 , `Dolomite_moisture`, `YPH_lag1`, `YPH_lag12`, and `Compost_lag1`.

C. Model Selection and Walk-Forward Validation

Six regression algorithms were evaluated: Ridge Regression, LASSO Regression, CatBoost, XGBoost, Random Forest, and a fully connected Neural Network. All models were assessed using an identical walk-forward validation protocol to ensure fair comparison under realistic forecasting conditions.

Walk-forward validation using a sliding window approach is a temporally ordered evaluation strategy in which the model is trained on a fixed-length historical window and tested on the immediately subsequent period, with the window advancing by one year at each step. This mirrors the real-world constraint that future yield cannot inform model training, and explicitly prevents the temporal data leakage that random splits introduce. In this implementation, a rolling 3-year training window was used, with evaluation beginning in 2020 using training data from 2017–2019 to predict 2020, and so forth through 2025.

The 3-year window length was selected after empirical comparison: windows shorter than 3 years caused underfitting due to insufficient training data, while windows longer than

3 years introduced noise from older, potentially unrepresentative estate conditions and reduced predictive stability. The 3-year window provided the most consistent generalization across all tested models. Performance was aggregated across all walk-forward test folds using root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), and the coefficient of determination (R^2).

D. Ridge Regression and Regularization

Ridge Regression minimizes the penalized residual sum of squares defined as:

$$\hat{\beta} = \arg \min \{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \} \quad (2)$$

where λ is the regularization hyperparameter tuned via cross-validation. The L_2 penalty distributes coefficient magnitudes across all correlated features rather than zeroing them out, which is critical in this context: fertilizer variables exhibit collinearity, and suppressing any of them (as LASSO does) would reduce interpretability and mask genuine agronomic contributions. The Ridge solution thus preserves all feature contributions at the expense of minor shrinkage, yielding a more stable and agronomically interpretable model [12].

E. Training and Testing Configuration

To evaluate the regression models under realistic deployment conditions, the 96-month dataset was processed through a rolling walk-forward validation framework. For each validation sequence, the target test year t was isolated chronologically within the range of 2020 to 2025, while the training data was restricted to a fixed 3-year historical block ($t-3$ to $t-1$). Consequently, each validation fold utilized exactly 36 monthly observations for training, while the remaining 12 months of the target year served as the unseen testing window.

Features were channeled through an automated execution pipeline where missing fields were resolved using median imputation and standardized via a standard scaling function. To prevent look-ahead bias, scaling parameters and medians were computed strictly on training data before being applied uniformly across all climate indicators and chemical variables. Model hyperparameters were held constant across all rolling iterations to ensure comparability. Ridge and LASSO regression employed regularization penalties of $\alpha = 5.0$ and $\alpha = 1.0$, respectively. Random Forest consisted of 300 trees with a maximum depth of 6 and a minimum leaf size of 5. XGBoost used 200 boosting rounds with a maximum depth of 3, a learning rate of 0.05, and row and column subsampling rates of 0.7. CatBoost was configured with 300 iterations, a depth of 3, a learning rate of 0.025, and an L_2 leaf regularization coefficient of 15. The neural network comprised two hidden layers with 16 and 8 ReLU units, a dropout rate of 0.2, He-normal initialization, and an Adam optimizer with a learning rate of 0.01, trained for 100 epochs using a batch size of 4.

Upon completion of walk-forward validation, a separate deployment model was trained on the most recent 3-year block (2023–2025) using the identical hyperparameters and preprocessing pipeline established through validation. This model

is intended for live 2026 yield forecasting within the estate management application and is not subject to retrospective evaluation, as no ground truth exists for the deployment period at the time of writing.

F. Explainable AI via Ridge Coefficient Analysis

Explainability is achieved through direct model-intrinsic analysis rather than post-hoc approximation methods such as SHAP or LIME. For each prediction, feature-level contributions are computed as:

$$c_i = x_i^{(\text{scaled})} \times \beta_i \quad (3)$$

where $x_i^{(\text{scaled})}$ is the standardized feature value and β_i is the corresponding Ridge coefficient. This produces a signed contribution score for each feature, indicating both the magnitude and direction of its influence on the current prediction. Features are ranked by absolute contribution and presented as an ordered explanation table. A natural language summary is then generated by passing this table to an LLM, producing human-readable agronomic explanations that estate managers and farmers can directly act upon.

This approach has important advantages over SHAP and LIME: it is computationally trivial, deterministic, requires no additional libraries, is always faithful to the actual model prediction, and produces contributions that sum exactly to the predicted value (after intercept adjustment). LIME, by contrast, trains a local surrogate model and introduces approximation error. Both SHAP and LIME treat the model as a black box, whereas Ridge coefficient analysis uses the model's own parameters directly, making it a true model-intrinsic XAI method.

IV. EXPERIMENTS AND RESULTS

A. Model Comparison

Table I presents the averaged performance metrics across all walk-forward test folds for each evaluated model. Ridge Regression achieves the best overall test R^2 of 0.9201 among the models that generalize well, with RMSE of 6.6283 and MAE of 5.5496. LASSO Regression achieves the lowest absolute error metrics numerically, but examination of its learned coefficients revealed that it concentrated importance almost entirely on `YPH_lag12`, effectively ignoring fertilizer and weather features. This represents a degenerate solution that discards agronomically meaningful variables and renders the model useless for providing actionable explanations about fertilizer or climate impacts on yield. Note that despite LASSO's numerically superior Test R^2 of 0.9687, this apparent accuracy is achieved at the cost of zeroing out 13 of 14 features, reducing it to a near-autoregressive predictor that provides no actionable information about fertilizer or weather impacts. Ridge avoids this behavior through its distributed L_2 penalty.

Tree-based ensemble methods (CatBoost, XGBoost, Random Forest) exhibit a larger gap between training and test performance, indicative of overfitting within the narrow 3-year window. Random Forest in particular collapses to a test R^2 of 0.3342, demonstrating that its high model capacity is

TABLE I
MODEL COMPARISON UNDER WALK-FORWARD VALIDATION (AVG.
ACROSS FOLDS)

Model	Test RMSE	Test MAE	Test MSE	Test R^2
CatBoost	10.073	7.059	103.68	0.8313
XGBoost	11.427	7.945	131.85	0.7855
Random Forest	20.209	15.273	409.56	0.3342
LASSO	3.036	2.527	18.99	0.9687*
Ridge	6.628	5.550	49.00	0.9201
Neural Net	10.678	9.165	120.28	0.8045

*LASSO achieves high R^2 by concentrating on YPH_lag12 and zeroing 13 of 14 features, yielding a degenerate autoregressive predictor unsuitable for agronomic decision support.

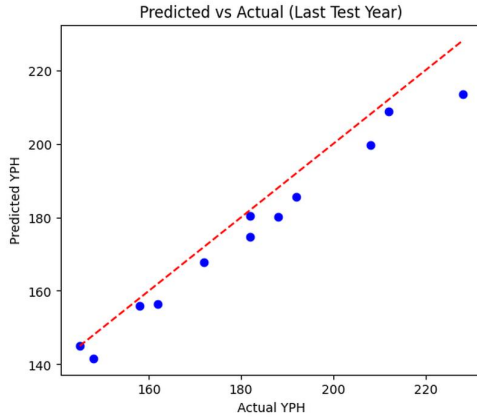


Fig. 1. Predicted vs. Actual YPH for the 2025 test year. Points close to the dashed identity line indicate accurate predictions. The model captures the range and trend of monthly yield with high fidelity.

unsuitable for the small-sample, temporally structured nature of estate yield data. The Neural Network achieves moderate test performance ($R^2 = 0.8045$) but with high variability across folds. Ridge Regression, with its closed-form solution and implicit regularization, is therefore the most appropriate model for this application domain.

B. Ridge Regression: 2025 Evaluation Fold

The most recent evaluable fold, trained on 2022–2024 and tested on 2025, serves as the representative performance snapshot, as it reflects the latest available ground truth and the most current estate conditions. This fold yields $R^2 = 0.9261$, RMSE = 6.23, and MAE = 5.41, confirming consistent performance with the averaged walk-forward metrics. The model captures over 92% of the variance in monthly tea yield. The slight under-prediction tendency observed in the residuals is a deliberate consequence of the Ridge regularization penalty, which biases the model toward conservative (lower-variance) predictions. This conservative bias is beneficial in the agricultural planning context: underestimating yield by a small margin is preferable to overestimating, as it avoids false confidence in resource allocation decisions. The Predicted vs. Actual and Residuals vs. Predicted plots for the 2025 test year are presented in Figs. 1 and 2.

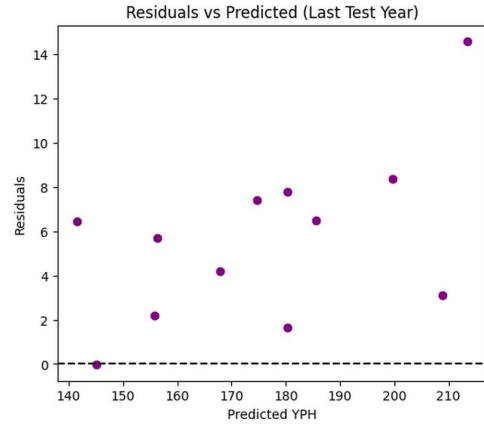


Fig. 2. Residuals vs. Predicted YPH for the 2025 test year. All residuals are positive, reflecting a consistent but small conservative under-prediction introduced by Ridge regularization. No systematic non-linear patterns are observed, confirming model adequacy.

C. Explainability Analysis

Feature contribution analysis from the Ridge coefficient model reveals consistent patterns across test folds. The lag features YPH_lag1 and YPH_lag12 are the strongest predictors, confirming the autocorrelated nature of tea yield and the importance of prior-month and same-month-prior-year yield as temporal anchors. Among weather variables, $rain_sum$ and ET_0 contribute significantly, reflecting the dependency of photosynthetic productivity and water stress on the rainfall-evapotranspiration balance. Among fertilizer variables, Dolomite and Urea consistently register positive contributions, consistent with agronomic literature on calcium-magnesium and nitrogen supplementation in tea cultivation. Importantly, the $Dolomite_moisture$ interaction feature, a novel derived variable introduced in this work, captures synergistic effects between Dolomite application and effective moisture availability that neither variable captures independently.

A critical distinction from prior work is that all 14 features retain non-zero contributions in the Ridge model, enabling the system to generate complete, agronomically rich explanations for each prediction. By contrast, in the LASSO model trained on the same data, 13 of the 14 features were zeroed out, leaving the only strongest lag variables active.

V. DISCUSSION

The results demonstrate that Ridge Regression with walk-forward validation outperforms all evaluated competitors in terms of balanced generalization: it avoids the degenerate feature collapse of LASSO, the overfitting of tree-based ensembles under small-sample temporal constraints, and the instability of Neural Networks across folds. The achieved test R^2 of 0.9201 compares favorably against Mampitiya et al. (2025) [4], who reported $R^2 \approx 0.9037$ using CatBoost with random splits. Critically, our evaluation is conducted under a stricter, temporally ordered protocol that prevents data leakage,

meaning the reported performance is a more conservative and reliable estimate of real-world forecasting accuracy.

The empirical window size analysis is a methodological contribution in its own right. The finding that 3 years optimally balances underfitting and noise introduction provides practitioners with a concrete, data-supported guideline for structuring temporal models in similarly sized tea estate datasets. Shorter windows (1–2 years) insufficiently captured seasonal cycles; longer windows (4–5 years) introduced older data reflecting different estate management practices, degrading model stability.

The XAI framework presents a practical advantage for deployment. Post-hoc methods like SHAP require the full training dataset to compute Shapley values and introduce approximation through sampling. LIME requires training a separate surrogate model per prediction. Both add computational overhead and opacity to the explanation process. The Ridge coefficient approach computes exact contributions in milliseconds using only the model parameters and the current input vector, making it ideally suited for real-time deployment in the Streamlit web application where estate managers require immediate, interpretable feedback.

Several limitations warrant acknowledgment. The proposed model is intentionally scoped to estate-level deployment. Due to significant variation in management practices and fertilizer regimes across Sri Lankan tea estates, cross-estate generalization is not claimed. This is consistent with precision agriculture principles, where site-specific models consistently outperform generic ones. The system is designed to be re-trained on locally collected data from any target estate, making it adaptable rather than universally generalizable. The model also currently relies on historical weather data rather than forecasted weather, meaning predictions require the current month's weather summary; integration with weather forecast APIs would enable true forward-looking yield forecasting. Finally, the omission of soil nutrient variables (e.g., nitrogen content) represents a gap identified by domain evaluators; future work should explore incorporating soil health indicators as additional features.

VI. CONCLUSION

This paper presented a temporally-aware, explainable machine learning system for tea yield prediction in Sri Lanka. By combining Ridge Regression with walk-forward validation, lag feature engineering, and model-intrinsic XAI through direct coefficient analysis, the system achieves a test R^2 of 0.9201 on real estate data while generating transparent, actionable explanations that practitioners can directly interpret. The use of walk-forward validation represents a methodological improvement over existing Sri Lankan tea yield research, providing more reliable and realistic performance estimates. The empirical selection of a 3-year rolling training window and the introduction of the Dolomite–moisture interaction feature constitute novel contributions to the domain.

The system is deployed as part of an integrated web application combining the yield prediction module with tea

leaf disease detection and a retrieval-augmented agricultural chatbot. Positive evaluations by domain experts from the Sri Lanka Tea Board and a regional tea estate confirm the practical relevance and readiness of the system for real-world adoption. Future work will explore retraining the model across multiple estates to assess cross-estate adaptability, with generalization expected only after site-specific calibration, integrate forward-looking weather forecasting APIs for proactive planning, and incorporate localized soil health metrics to further enhance agronomic accuracy.

ACKNOWLEDGMENT

The author sincerely thanks Dr. Prasan Yapa, Module Leader, and Ms. Ayesha Jayasundara, Lecturer at the Informatics Institute of Technology, for their invaluable academic guidance and support throughout this research as part of the GPTEA second-year project. The author gratefully acknowledges Ms. Sharubini Sathasivam for her vital assistance in manually collecting estate-level yield and fertilizer data from the tea estate, without which this study could not have been realized. Sincere gratitude is extended to Mr. D.M. Sriyan Disanayake, Administrative Officer at the tea estate, for providing access to historical estate records and offering invaluable domain expertise regarding tea plantation operations. The author thanks Mr. S. Karunakaran for facilitating the data collection process and coordinating access to estate records. The appreciation is also extended to Mr. Indika Hiran Wijesinghe, Deputy Director of IT at the Sri Lanka Tea Board, for his expert domain and technical evaluation of the system and for his constructive recommendations that strengthened the practical applicability of this research. The author further thanks Dr. Sahan Priyanayana for his thorough technical review of the system architecture and machine learning components. Finally, the author acknowledges the Informatics Institute of Technology and Robert Gordon University Aberdeen for providing the academic framework and institutional support for this research.

GENERATIVE AI DISCLOSURE STATEMENT

During the preparation of this work, the author used Claude (Anthropic) for linguistic and grammatical refinement of the manuscript. The author takes full responsibility for all intellectual content of this publication.

REFERENCES

- [1] Asian Dispatch, "Sri Lanka tea production down 15% in April over scorching sun," *Asian Dispatch*, May 26, 2024. [Online]. Available: <https://asiandispatch.net/global-tea-demand-booms-south-asian-workers-fight-climate-change/>. [Accessed: Jun. 19, 2026].
- [2] Central Bank of Sri Lanka, "Agriculture Sector Data Bulletin – December 2024," CBSL, Colombo, Sri Lanka, 2024. [Online]. Available: https://www.cbsl.gov.lk/sites/default/files/cbslweb_documents/statistics/agriculture_data_bulletin_2024_december_e.pdf. [Accessed: Jun. 19, 2026].
- [3] P. Weththasinghe and U. I. Disanayake, "Alternative and ICT-based advisory and extension approach in the tea sector in Sri Lanka: An explorative study," *Tropical Agricultural Research*, vol. 35, no. 1, pp. 46–60, 2024. <http://dx.doi.org/10.4038/tar.v35i1.8703>

- [4] L. Mampitiya, H. S. Sumanasekara, N. Rathnayake, Y. Hoshino, and U. Rathnayake, "Explainable artificial intelligence to estimate the Sri Lankan (Ceylon) tea crop yield," *Smart Agricultural Technology*, vol. 11, p. 100999, 2025. <http://dx.doi.org/10.1016/j.atech.2025.100999>
- [5] P. Deka, N. Tanti, and P. Neog, "Assessing predictive models for tea yield: A statistical and machine learning approach in Assam's Biswanath Chariali District," *Journal of Experimental Agriculture International*, vol. 46, no. 7, pp. 526–534, 2024. <http://dx.doi.org/10.9734/jeai/2024/v46i72605>
- [6] D. Batool et al., "A hybrid approach to tea crop yield prediction using simulation models and machine learning," *Plants*, vol. 11, no. 15, p. 1925, 2022. <http://dx.doi.org/10.3390/plants11151925>
- [7] P. Phan, N. Chen, L. Xu, and Z. Chen, "Using multi-temporal MODIS NDVI data to monitor tea status and forecast yield: A case study at Tanuyen, Laichau, Vietnam," *Remote Sensing*, vol. 12, no. 11, pp. 1814, 2020. <http://dx.doi.org/10.3390/rs12111814>
- [8] S. J. J. Jui et al., "Spatiotemporal hybrid random forest model for tea yield prediction using satellite-derived variables," *Remote Sensing*, vol. 14, no. 3, p. 805, 2022. <http://dx.doi.org/10.3390/rs14030805>
- [9] Q. Huang et al., "Automatic construction of knowledge graph of tea diseases and pests," *Annals of Computer Science and Information Systems*, vol. 37, pp. 141–146, 2023. <http://dx.doi.org/10.15439/2023F6100>
- [10] M. Aldawoud and Z. Sbaï , "Multi-level optimization-based ensemble machine learning for efficient crop yield prediction in Saudi Arabia," *Annals of Computer Science and Information Systems*, vol. 43, pp. 429–440, 2025. <http://dx.doi.org/10.15439/2025F5652>
- [11] Open-Meteo, "Open-Meteo historical weather archive API," 2024. [Online]. Available: <https://open-meteo.com/en/docs/historical-weather-api>. [Accessed: Jun. 19, 2026].
- [12] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. <http://dx.doi.org/10.2307/1267351>
- [13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. [Accessed: Jun. 19, 2026].