Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

# A Modular Agroclimatic Feature Engineering Framework for Country-Level Crop Yield Prediction Using XGBoost and LightGBM

**Ledyvia Audiz Coranov[1], Kusnawi[2]**

[1,2] Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

**Received:**
February 18, 2026
**Revised:**
July 19, 2026
**Accepted:**
August 4, 2026
**Published:**
August 29, 2026

Corresponding Author:
**Author Name*:**
Ledyvia Audiz Coranov
**Email*:**
ledyviaudiz@students.amikom.ac.id

DOI:
10.63158/journalisi.v8i4.1762



**Abstract.** Accurate crop yield prediction is essential for supporting food-security assessment, agricultural planning, and policy-level decision-making. This study proposes a modular agroclimatic feature engineering framework for country-level crop yield prediction using XGBoost and LightGBM. The proposed framework integrates climate indicators, climate–pesticide interactions, temporal descriptors, composite agroclimatic indices, and nonlinear transformations to improve predictive representation while controlling data leakage. Experiments were conducted on 28,151 country–crop–year observations from 98 countries covering 1990–2013. To evaluate temporal and geographic generalization, models were assessed using time-based validation, Random KFold, GroupKFold by country, bootstrap confidence intervals, and held-out-country evaluation. Results show that LightGBM with the S3 feature configuration achieved the best temporal prediction performance, obtaining an $R^2$ of 0.9492 and RMSE of 21,257 hg/ha. However, held-out-country evaluation revealed lower transferability, with the best configuration achieving $R^2$ of 0.6694, highlighting the challenge of geographic generalization. The findings demonstrate that engineered agroclimatic features can significantly improve country-level crop yield prediction, but model performance depends strongly on the validation setting. This framework provides a reliable benchmark for leakage-controlled agricultural machine learning research rather than direct farm-level operational forecasting.

**Keywords**: Agroclimatic feature engineering; Crop yield prediction; LightGBM; XGBoost; Geographic generalization

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

## 1. INTRODUCTION

Accurate and timely crop yield prediction supports food-security assessment, national production planning, and resource allocation. Machine-learning models can represent nonlinear associations among climate, agricultural inputs, and crop production that are difficult to capture with fixed-form statistical models[1], [2], [3], [4]. Gradient-boosting methods are particularly suitable for heterogeneous tabular data because they can model interactions without requiring a predetermined functional form [5], [6], [7].

Most crop-yield studies operate at field, farm, county, or regional scales, where detailed meteorological, soil, remote-sensing, and management observations are available [8], [9]. This study instead considers country-level historical records. Although national aggregation masks sub-national variability, such data remain useful for macro-level benchmarking because they provide broad temporal coverage, common crop categories, and comparable national reporting units. The framework is therefore intended for country-level forecasting research and policy-level analysis, not causal agronomic inference or farm-level operational recommendations.

Previous studies show that climate and agricultural inputs can improve yield prediction beyond historical production patterns [3], [4]. Feature engineering may further expose interactions, temporal dependencies, and nonlinear responses [8], [10]. However, its contribution is often evaluated under random partitions that allow repeated country-crop series to occur in both training and test sets. Such overlap can create optimistic performance estimates and obscure whether engineered features improve temporal or geographic transferability [11].

To address this limitation, the proposed framework separates five feature modules: agroclimatic indices (FC-1), climate-pesticide interactions (FC-2), temporal descriptors (FC-3), composite risk and potential indices (FC-4), and nonlinear transformations (FC-5). Their contributions are evaluated through cumulative scenarios S0-S4 and per-module ablations A1-A5 using XGBoost and LightGBM. The protocol combines a time-based split, Random KFold, GroupKFold by country, country-cluster bootstrap intervals, and a final held-out-country evaluation.

This study investigates three research questions: (RQ1) Which feature modules enhance country-level crop-yield prediction, and are their effects dependent on the model? (RQ2) How do model rankings vary across temporal and geographic validation protocols? (RQ3) To what extent do the selected configurations generalize to countries excluded prior to model development? The primary contribution is a validation-aware feature-scenario analysis that differentiates within-country temporal prediction from transfer to previously unseen countries.

## 2. METHODS

Figure 1 presents the leakage-controlled experimental workflow that structures the entire method. The process begins with 28,151 country–crop–year observations from 98 countries. Before feature-statistics fitting, categorical encoding, hyperparameter tuning, scenario selection, or model fitting, 19 countries were reserved once using seed 42, leaving 79 countries for development. Within the development countries, observations through 2009 were used for training and observations from 2010–2013 were used for the time-based test. The development data were then used to construct cumulative scenarios S0–S4 and per-module ablations A1–A5, tune XGBoost and LightGBM using 20 Optuna trials per configuration and select the final configurations.

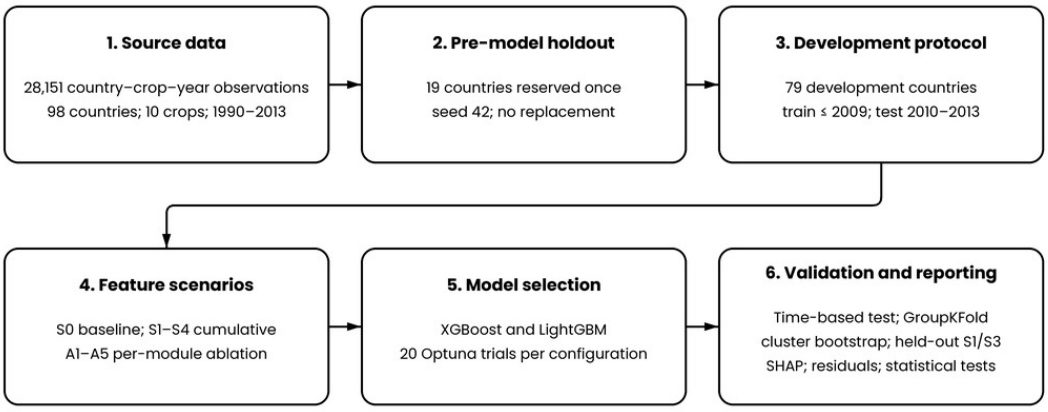


**Figure 1.** Leakage-controlled experimental workflow

As shown in Figure 1, model development and final geographic evaluation were deliberately separated. Random KFold and GroupKFold by country, country-cluster bootstrap intervals, statistical comparisons, SHAP analysis, and residual diagnostics were

produced within the development protocol. Only after tuning, scenario selection, and model selection had been completed were S1 and S3 evaluated on the same fixed held-out countries. This ordering prevents held-out-country outcomes from influencing model development. Sections 2.1–2.6 describe the problem formulation, data, leakage control, feature scenarios, model optimization, and evaluation procedures represented in the workflow.

### 2.1. Problem Formulation

Crop-yield prediction is formulated as supervised regression. For observation i, the input vector $x_i$ contains country and crop indicators, primitive climate and pesticide variables, and scenario-specific engineered features; $y_i$ is the reported crop yield in hectograms per hectare (hg/ha). The learning objective in equation (1) explicitly represents this regularized empirical risk objective.

$$\hat{f} = \arg\min_{f} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(y_i, f(x_i)\right) + \lambda \Omega(f) \right] \quad (1)$$

where n is the number of training observations, $f(x_i)$ is the prediction, $\mathcal{L}$ is the regression loss, $\Omega(f)$ is a model-complexity penalty, and $\lambda$ controls regularization. XGBoost and LightGBM are trained independently for each cumulative and ablation scenario. Equation (1) shows that model fitting balances predictive loss against model complexity, so scenario comparisons reflect both feature representation and regularized learner behavior.

### 2.2. Dataset

The source is the public Crop Yield Prediction Dataset distributed through Kaggle [12]. The Kaggle description attributes its components to international agricultural and development records; recent peer-reviewed global crop-data products similarly demonstrate the harmonization of national statistics and geospatial observations [13], [14]. The raw file contains 28,242 observations from 101 countries and ten crop types covering 1990-2013. After excluding three countries with fewer than 15 observed years, the final analytical dataset contains 28,151 country-crop-year observations from 98 countries; zero-pesticide records were retained as valid observations.

The recorded variables are country, crop type, year, annual rainfall, annual mean temperature, pesticide use, and crop yield. Rainfall and temperature values were taken directly from the source dataset and were not synthetically generated in this study. Annual Mauna Loa atmospheric $CO_2$ means were mapped by year from NOAA Global Monitoring Laboratory records [15]. Because $CO_2$ increases almost monotonically over the study period, it is interpreted as a temporal-trend proxy rather than independent causal evidence of fertilization. Figure 2 provides the descriptive context for the modeling task by summarizing the yield distribution, temporal trend, and crop-level differences in the analytical dataset. Figure 2 shows a strongly right-skewed yield distribution, a rising mean-yield trajectory over time, and substantial differences among crop categories. These patterns motivate nonlinear models and explicit controls for crop identity and temporal trend.

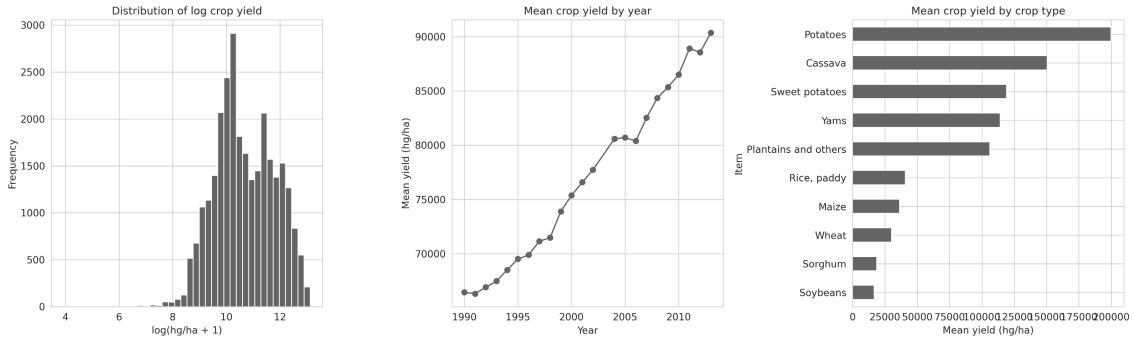


**Figure 2.** Exploratory summary of the crop-yield dataset

### 2.3. Preprocessing and Leakage Control

Nineteen of the 98 countries were selected once by simple random sampling without replacement using seed 42. The selection was not stratified by crop, yield, or climate. These countries were reserved before feature-statistics fitting, categorical encoding, hyperparameter tuning, scenario selection, and model fitting. The remaining 79 countries formed the development set. Crop coverage was retained across both groups, and the fixed held-out set was accessed only after development-stage selection was complete.

Country and crop type were one-hot encoded. The encoders were fitted only on development-training observations and used handle_unknown='ignore'. Consequently, all 4,298 rows from held-out countries were represented by zero across the 79 development-country indicator columns. Held-out predictions therefore relied on crop

identity, year, observed climate and pesticide variables, and scenario-specific engineered features rather than a country-specific indicator.

Within the development countries, the primary time-based split used observations through 2009 for training (19,657 rows) and 2010-2013 for testing (4,196 rows). A random 80:20 split was retained only as an optimistic benchmark. GroupKFold validation used country as the grouping variable. Feature-specific minimum, maximum, standard-deviation, and year-range statistics were fitted only on the relevant training partition and refitted within each GroupKFold training fold. Lag, rolling, difference, percentage-change, and cumulative variables were generated after sorting by country, crop, and year, and used only previous observations. Initial unavailable lags were set to zero.

### 2.4. Modular Agroclimatic Feature Engineering Framework

Five primitive numeric variables and 37 engineered variables were considered in addition to one-hot country and crop indicators. The modules contain FC-1 = 8, FC-2 = 6, FC-3 = 12, FC-4 = 4, and FC-5 = 7 features. Table 1 gives the exact definitions used in the notebook. Table 1 also links every computational definition to its intended predictive interpretation.

**Table 1.** Primitive and engineered variables used in the experiments

| Module | Code | Variable | Formula / description | Interpretation |
|---|---|---|---|---|
| Primitive | rainfall | Rainfall | Annual rainfall (mm) | Water availability |
| Primitive | temp_mean | Mean temperature | Annual mean temperature (°C) | Thermal conditions |
| Primitive | pesticide_use | Pesticide use | Annual pesticide use (tonnes) | Input intensity |
| Primitive | co2_ppm | Atmospheric $CO_2$ | Annual Mauna Loa $CO_2$ mapped by year | Temporal-trend proxy |
| Primitive | Year | Year | Observation year | Time control |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Module | Code | Variable | Formula / description | Interpretation |
|---|---|---|---|---|
| FC-1 | fc1_gdd | Thermal-excess proxy | max(temp_mean - 10, 0) | Annual thermal suitability |
| FC-1 | fc1_ombrothermic | Ombrothermic index | rainfall / (2 × \|temp_mean\|) | Humidity-temperature ratio |
| FC-1 | fc1_lang_rf | Lang rain factor | rainfall / \|temp_mean\| | Climate humidity ratio |
| FC-1 | fc1_htc | Hydrothermal coefficient | (rainfall × 10)/(fc1_gdd × 30) | Moisture-heat balance |
| FC-1 | fc1_aridity_dm | De Martonne index | rainfall/(temp_mean + 10) | Aridity indicator |
| FC-1 | fc1_pet | Temperature-based PET proxy | 16 × (10 × max(temp_mean,0)/120) ^1.514 | Potential water demand proxy |
| FC-1 | fc1_wbi | Water-balance index | rainfall - 12 × fc1_pet | Annual surplus/deficit proxy |
| FC-1 | fc1_wsr | Water-supply ratio | rainfall/ (12 × fc1_pet) | Water sufficiency proxy |
| FC-2 | fc2_pce | Pesticide per heat unit | pesticide_use/(fc1_gdd + 1) | Input relative to heat |
| FC-2 | fc2_pwi | Pesticide-water interaction | pesticide_use/(rainfall + 1) | Input relative to rainfall |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Module | Code | Variable | Formula / description | Interpretation |
|---|---|---|---|---|
| FC-2 | fc2_ocpi | Optimal climate-pesticide interaction | (opt_rain + opt_temp) × pest_norm | Climate-input interaction |
| FC-2 | fc2_pipu | Pesticide intensity per rainfall | pesticide_use/ (rainfall/100 + 1) | Alternative water scaling |
| FC-2 | fc2_tpe | Temperature-pesticide efficiency | temp_opt × pest_norm | Input under favorable temperature |
| FC-2 | fc2_cfwi | $CO_2$-water interaction | log1p(max(co2_ppm-354,0)) × clip(fc1_wsr,0,3) | Trend-water interaction |
| FC-3 | fc3_pest_lag1 | Pesticide lag 1 | shift (1) | One-year memory |
| FC-3 | fc3_pest_lag2 | Pesticide lag 2 | shift (2) | Two-year memory |
| FC-3 | fc3_pest_roll3 | Pesticide rolling mean | mean(shift(1).rolling(3)) | Past trend smoothing |
| FC-3 | fc3_pest_roll3_std | Pesticide rolling SD | std(shift(1).rolling(3)) | Past volatility |
| FC-3 | fc3_pest_delta | Past pesticide change | lag1 - lag2 | Input dynamics |
| FC-3 | fc3_pest_pct_change | Past percentage change | (lag1-lag2)/\|lag2\|, clipped [-2,2] | Relative dynamics |
| FC-3 | fc3_year_sin | Year sine | sin (2π × year_norm) | Smooth nonlinear |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Module | Code | Variable | Formula / description | Interpretation |
|---|---|---|---|---|
| | | | | time transform |
| FC-3 | fc3_year_cos | Year cosine | cos (2π × year_norm) | Smooth nonlinear time transform |
| FC-3 | fc3_decade_norm | Decade index | training-fitted normalized decade | Long-term period marker |
| FC-3 | fc3_rain_lag1 | Rainfall lag 1 | shift(1) | Past rainfall |
| FC-3 | fc3_rain_roll3 | Rainfall rolling mean | mean(shift(1).rolling(3)) | Past rainfall trend |
| FC-3 | fc3_pest_cumsum | Cumulative past pesticide | shift(1).cumsum() | Long-term prior exposure |
| FC-4 | fc4_cwpp | Water potential proportion | max(fc1_wbi,0)/(12 × fc1_pet + 1) | Water-surplus potential |
| FC-4 | fc4_tgsl | Thermal-suitability proxy | clip(fc1_gdd/max_train(fc1_gdd),0,1) | Normalized thermal suitability |
| FC-4 | fc4_yps | Yield-potential score | (rain_score + temp_score + pest_norm)/3 | Composite favorable conditions |
| FC-4 | fc4_ari | Agronomic-risk index | 0.5 × water_deficit + 0.5 × pesticide_volatility | Composite stress proxy |
| FC-5 | fc5_log_pest | Log pesticide | log1p(pesticide_use) | Variance stabilization |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Module | Code | Variable | Formula / description | Interpretation |
|---|---|---|---|---|
| FC-5 | fc5_log_rain | Log rainfall | log1p(rainfall) | Skewness reduction |
| FC-5 | fc5_sqrt_gdd | Square-root thermal proxy | sqrt(fc1_gdd) | Smoothed heat effect |
| FC-5 | fc5_rain_sq | Rainfall squared | $rainfall^2/10^6$ | Nonlinear rainfall effect |
| FC-5 | fc5_temp_sq | Temperature squared | $temp\_mean^2/100$ | Nonlinear temperature effect |
| FC-5 | fc5_gdd_rain | Thermal-rainfall interaction | fc1_gdd × rainfall/$10^5$ | Climate interaction |
| FC-5 | fc5_pest_year | Pesticide-year interaction | log1p(pesticide_use) × year_norm | Time-varying input intensity |

Table 1 confirms that the 37 engineered variables span climate ratios, input interactions, strictly lagged temporal descriptors, composite proxies, and nonlinear transforms. The De Martonne-type aridity variable is supported by recent bioclimatic applications [16], while the PET- and GDD-related features are treated only as data-limited proxies rather than causal physiological measurements. The standard daily growing-degree-day formulation and annual accumulation are given in Equations (2) and (3) [17]. Equation (2) defines daily thermal accumulation, whereas Equation (3) aggregates it over the days observed in year y.

$$\mathrm{GDD_d} = \max\left(\frac{\mathrm{T_{max,d}} + \mathrm{T_{min,d}}}{2} - \mathrm{T_{base}}, 0\right) \quad (2)$$

$$AGDD_y = \sum_{d=1}^{D_y} G\,DD_d \quad (3)$$

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Daily maximum and minimum temperatures are unavailable in the source data. The implemented variable is therefore the annual thermal-excess proxy in Equation (4), not accumulated daily GDD. Together, Equations (2) and (3) clarify the data requirements of the standard formulation and why it cannot be reproduced from annual mean temperature alone.

$$\mathrm{GDD}_{\mathrm{proxy,y}} = \max\left(\mathrm{T}_{\mathrm{mean,y}} - \mathrm{T}_{\mathrm{base}}, 0\right) \tag{4}$$

Equation (4) is therefore the implemented annual thermal-excess proxy; it preserves the threshold concept but does not estimate daily heat accumulation. The PET variable is likewise a temperature-based proxy inspired by Thornthwaite [18], not the full monthly Thornthwaite procedure. The year transformations use training-fitted bounds: Equations (5a)–(5c) then formalize the training-fitted year normalization and its two smooth nonlinear transformations.

$$\mathrm{Year}_{\mathrm{norm}} = \frac{\mathrm{Year} - \mathrm{Year}_{\mathrm{min,train}}}{\mathrm{Year}_{\mathrm{max,train}} - \mathrm{Year}_{\mathrm{min,train}}} \tag{5a}$$

$$\mathrm{Year}_{\sin} = \sin(2\pi \mathrm{Year}_{\mathrm{norm}}) \tag{5b}$$

$$\mathrm{Year}_{\cos} = \cos(2\pi \mathrm{Year}_{\mathrm{norm}}) \tag{5c}$$

Because the study period is not a repeating seasonal cycle, the sine and cosine terms are interpreted as smooth nonlinear time transforms. All module definitions are empirical transformations and should not be interpreted as independently validated causal agronomic mechanisms. Equation (5a) prevents test-period bounds from entering preprocessing, while Equations (5b) and (5c) provide bounded nonlinear representations of long-term time. Table 2 translates the modular definitions into the exact cumulative and single-module experimental scenarios used for model comparison.

**Table 2.** Feature composition of cumulative and ablation scenarios

| Scenario | Feature composition | Count |
|---|---|---|
| S0 - Baseline | Country one-hot (79) + crop one-hot (10) + Year | 90 |
| S1 - Primitive | S0 + rainfall, temp_mean, pesticide_use, co2_ppm | 94 |
| S2 - FC1+FC2 | S1 + FC-1 + FC-2 | 108 |
| S3 - FC1+FC2+FC3 | S2 + FC-3 | 120 |
| S4 - Full FC | S3 + FC-4 + FC-5 | 131 |
| A1-A5 | S1 + exactly one of FC-1, FC-2, FC-3, FC-4, or FC-5 | 98-106 |

Table 2 shows that model dimensionality increases from 90 predictors in S0 to 131 in S4, while A1–A5 isolate one module at a time; this separation is essential for distinguishing cumulative interaction effects from individual module contributions.

### 2.5. Model Training and Hyperparameter Optimization

XGBoost and LightGBM were evaluated because both are strong gradient-boosted decision-tree implementations for tabular data [5], [6]. XGBoost regularizes each tree according to the explicit XGBoost tree complexity term given in Equation (6).

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_{kj}^2 \quad (6)$$

where $T_k$ is the number of leaves, $w_{kj}$ is the weight of leaf j, $\gamma$ penalizes tree complexity, and $\lambda$ applies L2 regularization. LightGBM uses histogram-based, leaf-wise growth and its own regularization parameters [6]. Equation (6) shows that additional leaves and large leaf weights are penalized, thereby limiting overly complex trees; LightGBM controls analogous behavior through its own leaf-wise architecture and regularization settings[6]. Each of the ten scenarios was tuned independently for both models using Optuna with TPESampler(seed=42) [19]. Twenty trials were allocated to every model-scenario combination, producing 400 trials. This fixed budget was selected to keep the ten-scenario, two-model comparison computationally feasible and strictly comparable; it is a bounded search rather than a claim of exhaustive optimization. The objective maximized mean three-fold GroupKFold $R^2$ minus 0.5 times its standard deviation, thereby rewarding both accuracy and stability across countries. Feature statistics required by FC-2 and FC-4 were refitted inside each tuning fold. Table 3 reports the complete and identical search domains applied to each model–scenario combination.

**Table 3.** Optuna search spaces used for independent scenario tuning.

| Model | Parameters and ranges |
|---|---|
| XGBoost | n_estimators: 100-350; max_depth: 3-7; learning_rate: 0.01-0.3 (log); subsample: 0.6-1; colsample_bytree: 0.5-1; min_child_weight: 1-10; gamma: 0-1; reg_alpha: 0-2; reg_lambda: 0-2 |
| LightGBM | n_estimators: 100-350; max_depth: 3-7; learning_rate: 0.01-0.3 (log); num_leaves: 15-63; subsample: 0.6-1; colsample_bytree: 0.5-1; min_child_samples: 5-40; reg_alpha: 0-2; reg_lambda: 0-2 |

Table 3 demonstrates that both model families were tuned under bounded but comparable ranges; therefore, observed performance differences are less likely to arise from unequal search budgets. Experiments were run in Google Colab on Linux with 4 physical and 8 logical CPU cores, 50.99 GB RAM, and no GPU. Software versions were Python 3.12.13, NumPy 2.0.2, pandas 2.2.2, scikit-learn 1.6.1, XGBoost 3.3.0, LightGBM 4.6.0, Optuna 4.9.0, SHAP 0.52.0, SciPy 1.16.3. The complete pipeline required 22.2 minutes, including tuning; post-tuning model training required 6.0 minutes.

### 2.6. Evaluation and Statistical Analysis

Performance was reported using $R^2$, RMSE, MAE, and MAPE. Model selection emphasized $R^2$, RMSE, and MAE. MAPE was retained as a complementary relative-error measure because low-yield observations can generate disproportionately large percentage errors. WMAPE and sMAPE were therefore calculated for the selected time-based configuration. Random KFold and GroupKFold by country used five folds. Time-based and held-out-country uncertainty was quantified using 2,000 country-cluster bootstrap resamples, preserving within-country dependence. Paired model and scenario comparisons used country-level MAE as the statistical unit. Both paired t-tests with Cohen's d_z and Wilcoxon signed-rank tests with rank-biserial effect sizes were reported. The Friedman test compared eight model-scenario combinations, followed by a Nemenyi post-hoc procedure [20], [21], [22]. Residual normality was assessed separately as a diagnostic and not used as the sampling unit for paired tests. Interpretability analysis used the final time-based configuration, S3-LightGBM, and included feature importance, SHAP values [23], redundancy diagnostics, VIF with an intercept [24], and residual analysis. Because several engineered variables are deterministic transformations of others, infinite or very large VIF values are treated as evidence of redundancy rather than stable coefficients for individual attribution.

## 3. RESULTS AND DISCUSSION

### 3.1. Cumulative Scenario Performance

Table 4 reports both the optimistic random split and the primary time-based test. The final development-stage selection was based on time-based $R^2$, with RMSE and MAE used as supporting criteria. The table directly contrasts overlap-prone random partitioning with forward-looking temporal testing for every cumulative scenario.

**Table 4.** Performance across cumulative scenarios under random and time-based splits

| Split | Scenario | Model | $R^2$ | RMSE | MAE | MAPE (%) |
|---|---|---|---|---|---|---|
| Random | S0 - Baseline | XGBoost | 0.9567 | 17,390 | 11,043 | 30.28 |
| Random | S0 - Baseline | LightGBM | 0.8648 | 30,724 | 16,941 | 45.02 |
| Time-based | S0 - Baseline | XGBoost | 0.9184 | 26,938 | 16,408 | 34.59 |
| Time-based | S0 - Baseline | LightGBM | 0.8722 | 33,714 | 20,314 | 44.09 |
| Random | S1 - Primitive | XGBoost | 0.9470 | 19,246 | 12,044 | 32.74 |
| Random | S1 - Primitive | LightGBM | 0.9185 | 23,851 | 14,077 | 37.24 |
| Time-based | S1 - Primitive | XGBoost | 0.9217 | 26,382 | 16,224 | 35.10 |
| Time-based | S1 - Primitive | LightGBM | 0.9045 | 29,141 | 17,657 | 38.20 |
| Random | S2 - FC1+FC2 | XGBoost | 0.9680 | 14,958 | 8,887 | 23.67 |
| Random | S2 - FC1+FC2 | LightGBM | 0.9328 | 21,668 | 12,577 | 34.22 |
| Time-based | S2 - FC1+FC2 | XGBoost | 0.9411 | 22,885 | 13,312 | 28.10 |
| Time-based | S2 - FC1+FC2 | LightGBM | 0.9142 | 27,621 | 16,424 | 35.79 |
| Random | S3 - FC1+FC2+FC3 | XGBoost | 0.9697 | 14,536 | 8,958 | 24.20 |
| Random | S3 - FC1+FC2+FC3 | LightGBM | 0.9772 | 12,607 | 7,094 | 18.38 |
| Time-based | S3 - FC1+FC2+FC3 | XGBoost | 0.9392 | 23,251 | 13,462 | 27.49 |
| Time-based | S3 - FC1+FC2+FC3 | LightGBM | 0.9492 | 21,257 | 11,929 | 23.37 |
| Random | S4 - Full FC | XGBoost | 0.9672 | 15,143 | 9,163 | 24.64 |
| Random | S4 - Full FC | LightGBM | 0.9536 | 17,998 | 11,123 | 30.86 |
| Time-based | S4 - Full FC | XGBoost | 0.9388 | 23,336 | 13,551 | 28.64 |
| Time-based | S4 - Full FC | LightGBM | 0.9240 | 25,997 | 15,432 | 32.80 |

The best time-based configuration was S3-LightGBM ($R^2$ = 0.9492, RMSE = 21,257 hg/ha, MAE = 11,929 hg/ha). S2-XGBoost ranked next by $R^2$ (0.9411). The winner therefore differs from the earlier manuscript version: temporal descriptors improved LightGBM only when combined with FC-1 and FC-2. Random-split performance was consistently more optimistic and reached 0.9772 for S3-LightGBM, reinforcing that repeated country-crop histories should not be randomly divided when generalization is the target. Table 4 therefore identifies S3–LightGBM as the temporal winner and quantifies the optimism introduced by random splitting, particularly for recurrent country–crop histories. Figure 3 visualizes the validation-dependent changes in $R^2$ and RMSE across S0–S4 for both algorithms.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

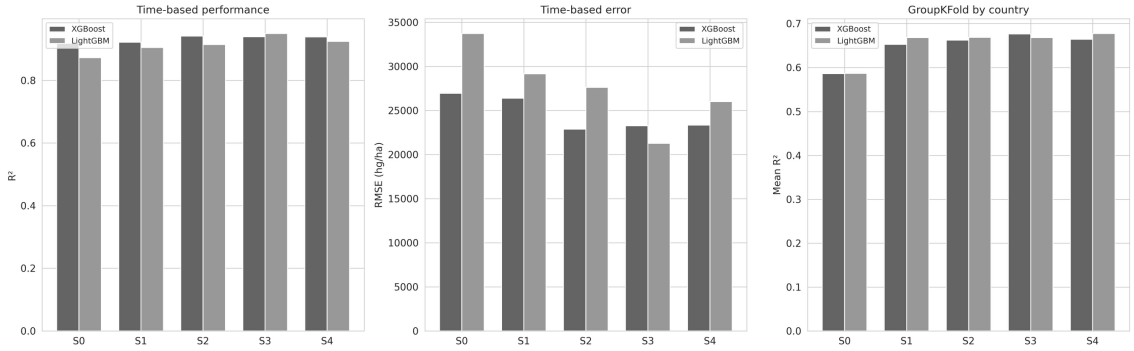


**Figure 3.** Validation-dependent comparison of cumulative scenarios

Figure 3 makes the ranking shift visible: performance improves sharply under random evaluation, while country-grouped scores remain much lower and comparatively compressed. Figure 4 compares observed and predicted yields for both algorithms under S3 on the time-based test set.

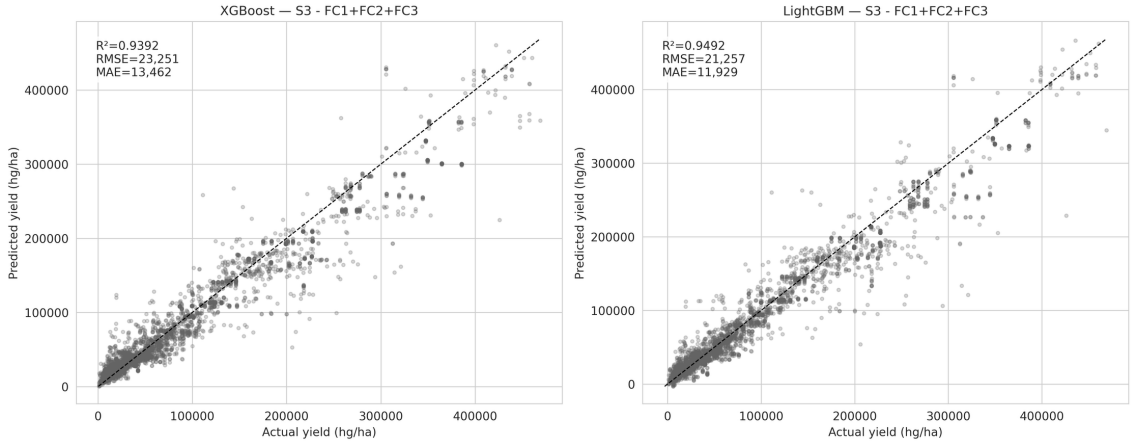


**Figure 4.** Actual versus predicted yield for XGBoost and LightGBM under S3

Figure 4 shows that both models follow the 1:1 line across most of the yield range. However, prediction dispersion increases at higher observed yields, and the tighter LightGBM fit is consistent with its lower RMSE and MAE under S3.

### 3.2. Per-Module Ablation

A1-A5 add one module at a time to S1, separating individual module effects from cumulative interactions. Table 5 quantifies each isolated module's incremental effect relative to the primitive S1 benchmark.

**Table 5.** Per-module ablation under the time-based split, relative to S1

| Scenario | Model | $R^2$ | $\Delta R^2$ vs. S1 | Effect |
|---|---|---|---|---|
| A1 - Primitive+FC1 | XGBoost | 0.8941 | -0.0276 | Hurts |
| A1 - Primitive+FC1 | LightGBM | 0.8946 | -0.0099 | Hurts |
| A2 - Primitive+FC2 | XGBoost | 0.9217 | +0.0000 | Neutral |
| A2 - Primitive+FC2 | LightGBM | 0.9357 | +0.0312 | Helps |
| A3 - Primitive+FC3 | XGBoost | 0.9425 | +0.0207 | Helps |
| A3 - Primitive+FC3 | LightGBM | 0.9157 | +0.0112 | Helps |
| A4 - Primitive+FC4 | XGBoost | 0.9068 | -0.0149 | Hurts |
| A4 - Primitive+FC4 | LightGBM | 0.9279 | +0.0234 | Helps |
| A5 - Primitive+FC5 | XGBoost | 0.9510 | +0.0293 | Helps |
| A5 - Primitive+FC5 | LightGBM | 0.9146 | +0.0101 | Helps |

The effects were model-dependent. FC-1 reduced $R^2$ for both models. FC-3 and FC-5 produced the largest isolated gains for XGBoost (+0.0207 and +0.0293), whereas FC-2 and FC-4 produced the largest gains for LightGBM (+0.0312 and +0.0234). This pattern explains why no single module can be described as universally beneficial. Table 5 confirms that module utility is neither monotonic nor algorithm-independent, which rules out interpreting feature engineering as a uniformly beneficial preprocessing step.

### 3.3. Cross-Validation and Bootstrap Uncertainty

Table 6 compares five-fold Random KFold and country-grouped cross-validation to assess how geographic separation changes average performance and uncertainty. The reported 95% confidence intervals were calculated from fold-to-fold variation using a student's t-distribution with five folds. These intervals represent uncertainty across cross-validation folds and should be distinguished from the 2,000-resample country-cluster bootstrap intervals used for the time-based and held-out-country evaluations in Tables 7 and 13.

**Table 6.** Five-fold Random KFold and GroupKFold-by-country results for all scenarios

| CV protocol | Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE | MAPE (%) |
|---|---|---|---|---|---|---|
| Random KFold | S0 - Baseline | XGBoost | 0.9555 [0.9524, 0.9586] | 17,475 [16,733, 18,216] | 11,276 | 31.76 |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| CV protocol | Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE | MAPE (%) |
|---|---|---|---|---|---|---|
| Random KFold | S0 - Baseline | LightGBM | 0.8827 [0.8678, 0.8976] | 28,361 [26,109, 30,612] | 16,357 | 45.17 |
| GroupKFold | S0 - Baseline | XGBoost | 0.5864 [0.4374, 0.7353] | 52,834 [35,267, 70,400] | 30,676 | 64.21 |
| GroupKFold | S0 - Baseline | LightGBM | 0.5869 [0.4365, 0.7373] | 52,839 [34,894, 70,784] | 30,880 | 68.94 |
| Random KFold | S1 - Primitive | XGBoost | 0.9479 [0.9449, 0.9508] | 18,917 [17,977, 19,856] | 11,828 | 32.07 |
| Random KFold | S1 - Primitive | LightGBM | 0.9231 [0.9170, 0.9292] | 22,975 [21,564, 24,387] | 13,862 | 36.44 |
| GroupKFold | S1 - Primitive | XGBoost | 0.6528 [0.5779, 0.7277] | 48,362 [35,590, 61,135] | 28,707 | 63.65 |
| GroupKFold | S1 - Primitive | LightGBM | 0.6681 [0.5752, 0.7610] | 47,101 [34,130, 60,073] | 28,179 | 61.09 |
| Random KFold | S2 - FC1+FC2 | XGBoost | 0.9668 [0.9637, 0.9698] | 15,092 [14,272, 15,913] | 8,781 | 23.31 |
| Random KFold | S2 - FC1+FC2 | LightGBM | 0.9348 [0.9301, 0.9394] | 21,163 [19,934, 22,391] | 12,512 | 34.25 |
| GroupKFold | S2 - FC1+FC2 | XGBoost | 0.6626 [0.5822, 0.7429] | 47,656 [34,759, 60,553] | 28,232 | 57.84 |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| CV protocol | Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE | MAPE (%) |
|---|---|---|---|---|---|---|
| GroupKFold | S2 - FC1+FC2 | LightGBM | 0.6689 [0.5788, 0.7589] | 47,079 [33,919, 60,240] | 27,817 | 59.24 |
| Random KFold | S3 - FC1+FC2+FC3 | XGBoost | 0.9692 [0.9669, 0.9714] | 14,542 [13,733, 15,350] | 8,855 | 23.91 |
| Random KFold | S3 - FC1+FC2+FC3 | LightGBM | 0.9774 [0.9761, 0.9787] | 12,460 [11,883, 13,037] | 6,947 | 18.24 |
| GroupKFold | S3 - FC1+FC2+FC3 | XGBoost | 0.6765 [0.5926, 0.7604] | 46,414 [34,551, 58,278] | 27,760 | 57.51 |
| GroupKFold | S3 - FC1+FC2+FC3 | LightGBM | 0.6685 [0.5909, 0.7461] | 46,901 [36,185, 57,618] | 28,228 | 58.47 |
| Random KFold | S4 - Full FC | XGBoost | 0.9658 [0.9632, 0.9685] | 15,314 [14,508, 16,120] | 9,149 | 24.78 |
| Random KFold | S4 - Full FC | LightGBM | 0.9528 [0.9501, 0.9555] | 17,998 [17,150, 18,845] | 11,065 | 30.33 |
| GroupKFold | S4 - Full FC | XGBoost | 0.6644 [0.5594, 0.7694] | 47,157 [35,027, 59,288] | 27,606 | 55.05 |
| GroupKFold | S4 - Full FC | LightGBM | 0.6774 [0.5718, 0.7830] | 46,381 [32,970, 59,792] | 27,101 | 57.27 |

Random KFold produced $R^2$ values of 0.883-0.977, whereas GroupKFold by country produced 0.586-0.677. The highest individual GroupKFold value was S4-LightGBM (0.6774), but S3 had the highest mean $R^2$ across both models (0.6725). Thus, temporal and geographic rankings are related but not identical. The wide GroupKFold intervals in Table

6 indicate substantial between-country heterogeneity, whereas the narrow Random KFold intervals largely reflect repeated geographic patterns shared across folds. Table 7 complements the fold-based analysis by reporting country-cluster bootstrap intervals for every time-based model–scenario result.

**Table 7.** Country-cluster bootstrap intervals for the time-based test set

| Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE [95% CI] | MAPE [95% CI] |
|---|---|---|---|---|---|
| S0 - Baseline | XGBoost | 0.9184 [0.8819, 0.9356] | 26,938 [23,653, 30,250] | 16,408 [14,994, 18,241] | 34.59 [28.45, 43.59] |
| S0 - Baseline | LightGBM | 0.8722 [0.7853, 0.9122] | 33,714 [29,348, 40,417] | 20,314 [17,622, 24,944] | 44.09 [34.44, 59.57] |
| S1 - Primitive | XGBoost | 0.9217 [0.8733, 0.9447] | 26,382 [23,382, 30,608] | 16,224 [14,341, 19,053] | 35.10 [28.30, 45.10] |
| S1 - Primitive | LightGBM | 0.9045 [0.8496, 0.9295] | 29,141 [25,945, 33,673] | 17,657 [15,515, 20,923] | 38.20 [30.70, 49.88] |
| S2 - FC1+FC2 | XGBoost | 0.9411 [0.9082, 0.9565] | 22,885 [20,202, 26,420] | 13,312 [11,672, 15,812] | 28.10 [21.90, 37.37] |
| S2 - FC1+FC2 | LightGBM | 0.9142 [0.8629, 0.9380] | 27,621 [24,396, 32,049] | 16,424 [14,352, 19,617] | 35.79 [28.73, 46.61] |
| S3 - FC1+FC2+FC3 | XGBoost | 0.9392 [0.9100, 0.9528] | 23,251 [20,412, 26,584] | 13,462 [11,889, 15,783] | 27.49 [22.27, 35.07] |
| S3 - FC1+FC2+FC3 | LightGBM | 0.9492 [0.9191, 0.9642] | 21,257 [18,316, 25,130] | 11,929 [10,433, 14,115] | 23.37 [18.37, 30.60] |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE [95% CI] | MAPE [95% CI] |
|---|---|---|---|---|---|
| S4 - Full FC | XGBoost | 0.9388 [0.8994, 0.9592] | 23,336 [19,846, 27,730] | 13,551 [11,596, 16,421] | 28.64 [22.69, 37.57] |
| S4 - Full FC | LightGBM | 0.9240 [0.8766, 0.9470] | 25,997 [22,760, 30,472] | 15,432 [13,301, 18,555] | 32.80 [26.10, 42.94] |

Table 7 shows that S3–LightGBM retains the highest time-based point estimate and favorable error intervals, although overlapping intervals caution against interpreting small numerical differences as universal superiority.

### 3.4. Statistical Comparisons

Model comparisons were based on paired country-level MAE. The difference was defined as XGBoost MAE minus LightGBM MAE; positive values favor LightGBM and negative values favor XGBoost. Table 8 reports both significance and effect-size evidence for the paired algorithm comparison within each cumulative scenario.

**Table 8.** XGBoost versus LightGBM under cumulative engineered-feature scenarios.

| Scenario | Lower country MAE | Paired t-test | Cohen d_z | Wilcoxon p | Rank-biserial r |
|---|---|---|---|---|---|
| S1 - Primitive | XGBoost | t(78)=-4.919; p <0.0001 | -0.553 | <0.0001 | -0.618 |
| S2 - FC1+FC2 | XGBoost | t(78)=-6.669; p <0.0001 | -0.750 | <0.0001 | -0.814 |
| S3 - FC1+FC2+FC3 | LightGBM | t(78)=3.721; p 0.0004 | 0.419 | 0.0002 | 0.487 |
| S4 - Full FC | XGBoost | t(78)=-7.315; p <0.0001 | -0.823 | <0.0001 | -0.828 |

XGBoost had lower country-level MAE under S1, S2, and S4, whereas LightGBM was significantly better under S3 ($d_z = 0.419$, $p = 0.0004$). This result supports validation- and configuration-specific model selection rather than a universal algorithm ranking. Table 8

demonstrates a genuine reversal under S3: LightGBM is favored only for that configuration, whereas XGBoost has lower country-level MAE in S1, S2, and S4. Table 9 extends the analysis from algorithm contrasts to within-model scenario changes, using S0 or S1 as the paired reference.

**Table 9.** Effect sizes for key cumulative scenario comparisons using country-level MAE

| Model | Comparison | Mean MAE difference | d_z | t-test p | Rank-biserial r | Wilcoxon p |
|---|---|---|---|---|---|---|
| XGBoost | S0 - Baseline → S1 - Primitive | -1,299 | -0.296 | 0.0103 | -0.333 | 0.0102 |
| XGBoost | S1 - Primitive → S2 - FC1+FC2 | 3,165 | 0.880 | <0.0001 | 0.813 | <0.0001 |
| XGBoost | S1 - Primitive → S3 - FC1+FC2+FC3 | 3,669 | 0.897 | <0.0001 | 0.892 | <0.0001 |
| XGBoost | S1 - Primitive → S4 - Full FC | 2,893 | 0.977 | <0.0001 | 0.871 | <0.0001 |
| LightGBM | S0 - Baseline → S1 - Primitive | 6,296 | 0.513 | <0.0001 | 0.641 | <0.0001 |
| LightGBM | S1 - Primitive → S2 - FC1+FC2 | 1,274 | 0.348 | 0.0028 | 0.460 | 0.0004 |
| LightGBM | S1 - Primitive → S3 - FC1+FC2+FC3 | 7,968 | 1.060 | <0.0001 | 0.941 | <0.0001 |
| LightGBM | S1 - Primitive → S4 - Full FC | 3,505 | 0.652 | <0.0001 | 0.737 | <0.0001 |

Table 9 indicates that the largest scenario-level improvement is obtained for S1→S3 in LightGBM (d_z = 1.060), while the negative S0→S1 effect for XGBoost shows that adding primitive variables does not automatically improve every learner. The Friedman test was significant ($\chi^2$ = 236.603, p = 1.96e-47, n = 79 countries). The Nemenyi procedure identified 17 significant pairs among 28 comparisons, with a critical difference of 1.181.

### 3.5. Model Interpretation, Redundancy, and Residual Diagnostics

Figure 5 compares the top-ten split-based feature-importance rankings for XGBoost and LightGBM under the selected S3 representation. Figure 5 reveals different internal emphases: XGBoost assigns substantial importance to country and crop indicators together with engineered indices, whereas LightGBM places stronger weight on $CO_2$, pesticide use, temperature, rainfall, and temporal descriptors. Figures 6 and 7 use SHAP summaries to examine both global importance and the direction and distribution of feature contributions for XGBoost and LightGBM, respectively.

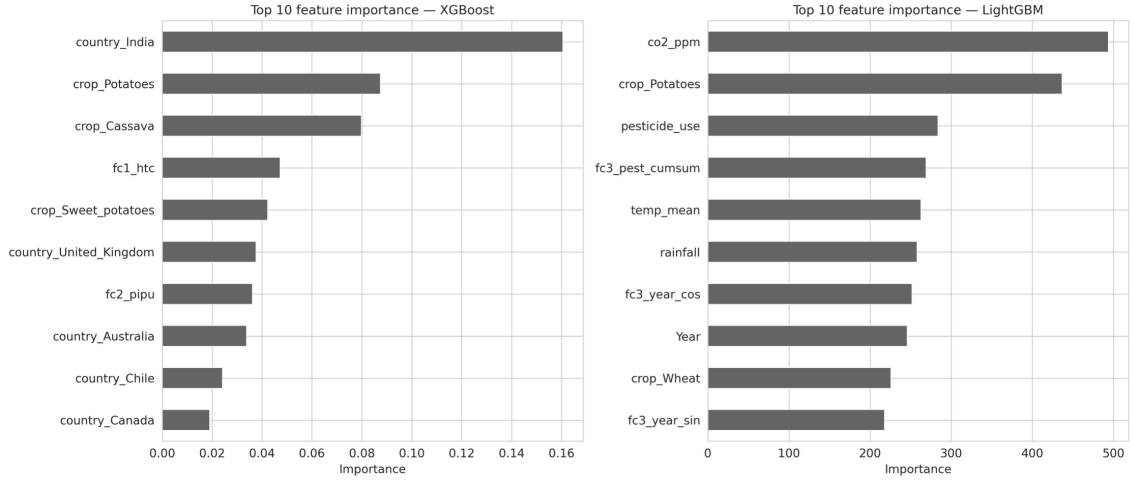


**Figure 5.** Top-ten feature importance for XGBoost and LightGBM under S3

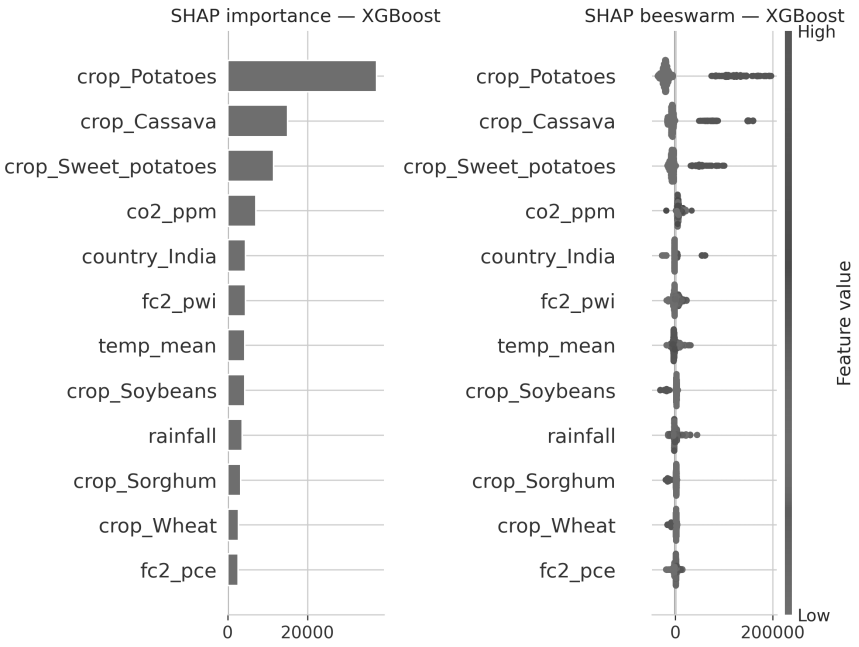


**Figure 6.** SHAP summary for XGBoost under S3

Crop identity dominates the SHAP rankings, particularly potatoes, cassava, soybeans, wheat, sorghum, sweet potatoes, maize, and rice. $CO_2$, rainfall, pesticide use, the cumulative past-pesticide feature, and year-related transforms also influence predictions. These values explain the fitted models' behavior; they do not establish agronomic causation. The prominence of crop and country indicators is consistent with characteristic national and crop-specific yield distributions. Specifically, Figure 6 shows a more concentrated XGBoost attribution pattern dominated by potatoes and cassava, whereas Figure 7 displays a broader LightGBM allocation across crop identities and primitive and temporal variables.

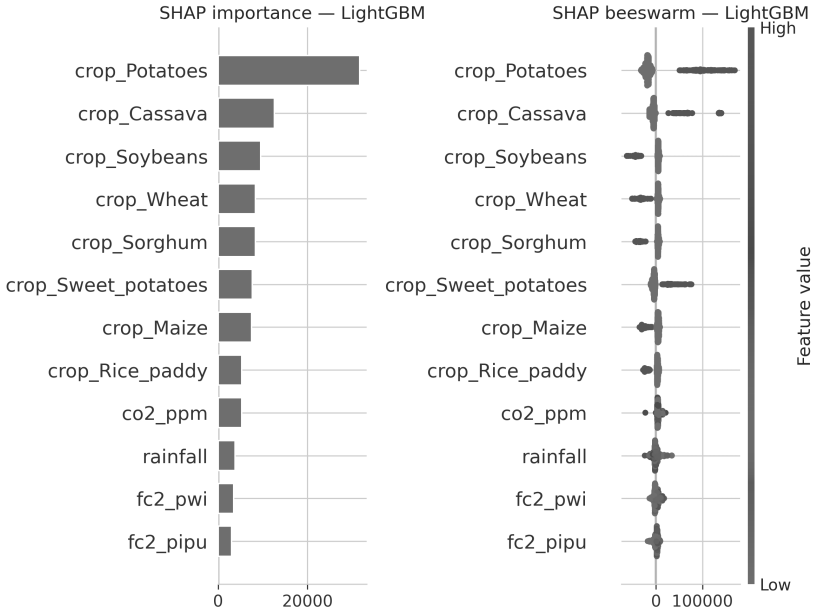


**Figure 7.** SHAP summary for LightGBM under S3

Figure 8 evaluates multicollinearity among the continuous S3 predictors using VIF calculated with an intercept. The corrected VIF analysis identified exact or near-exact dependencies. Ombrothermic and Lang ratios, several rainfall summaries, and PET-related transforms produced infinite or extremely large VIF values. Year and $CO_2$ also remained highly collinear (VIF approximately 799 and 769) because $CO_2$ was mapped from year. Tree models can still predict under such redundancy, but separate SHAP attributions for correlated variables are not uniquely identifiable and should be interpreted jointly [24]. The extreme values in Figure 8 should therefore be read as structural redundancy diagnostics, not as evidence that tree-based predictive accuracy is invalid. Table 10

identifies representative S4 feature pairs whose absolute Pearson correlations approach unity.

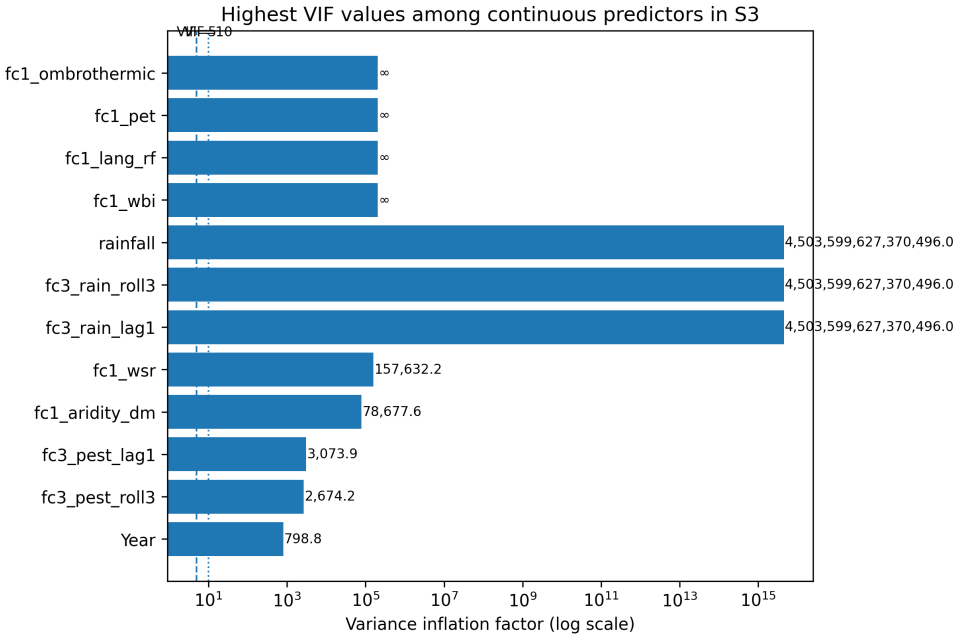


**Figure 8.** Corrected VIF diagnostic for continuous S3 predictors, calculated with an intercept

**Table 10.** Representative highly correlated feature pairs in the full S4 feature set

| Feature A | Feature B | \|Pearson r\| |
|---|---|---|
| fc1_gdd | fc4_tgsl | 1.0000 |
| fc3_rain_lag1 | fc3_rain_roll3 | 1.0000 |
| fc1_ombrothermic | fc1_lang_rf | 1.0000 |
| fc3_pest_lag1 | fc3_pest_roll3 | 0.9988 |
| fc1_wsr | fc4_cwpp | 0.9974 |
| fc1_pet | fc4_tgsl | 0.9972 |
| fc1_gdd | fc1_pet | 0.9972 |
| fc3_pest_lag2 | fc3_pest_roll3 | 0.9971 |

Table 10 confirms that several modules reproduce nearly identical information through rescaling, rolling summaries, or deterministic transformations; this redundancy provides a concrete explanation for the non-additive behavior of S4. Table 11 localizes the largest S3–LightGBM errors by country, year, and crop rather than relying only on aggregate metrics.

**Table 11.** Highest residual errors for the selected S3-LightGBM time-based model

| Group type | Group | n | MAE | RMSE | MAPE (%) | Mean residual |
|---|---|---|---|---|---|---|
| Country | Guyana | 24 | 47,110 | 76,905 | 57.41 | 12,991 |
| Country | Portugal | 24 | 35,173 | 51,957 | 40.06 | 25,236 |
| Country | Malaysia | 16 | 33,593 | 43,635 | 32.56 | 10,849 |
| Year | 2013 | 1049 | 14,020 | 24,167 | 27.11 | 6,798 |
| Year | 2012 | 1049 | 12,727 | 23,605 | 23.96 | 3,488 |
| Year | 2011 | 1049 | 11,474 | 20,063 | 21.85 | 3,639 |
| Crop | Cassava | 328 | 24,625 | 31,794 | 22.41 | 9,744 |
| Crop | Potatoes | 636 | 22,675 | 31,672 | 14.57 | 9,314 |
| Crop | Sweet potatoes | 412 | 21,377 | 36,638 | 19.70 | 10,610 |

Table 11 shows that high-error groups are concentrated rather than uniform: Guyana, the final test year, and high-yield root crops contribute disproportionately to the overall error profile. Figure 9 provides complementary residual-versus-fitted, distributional, Q–Q, and scale-location diagnostics for the same selected model.

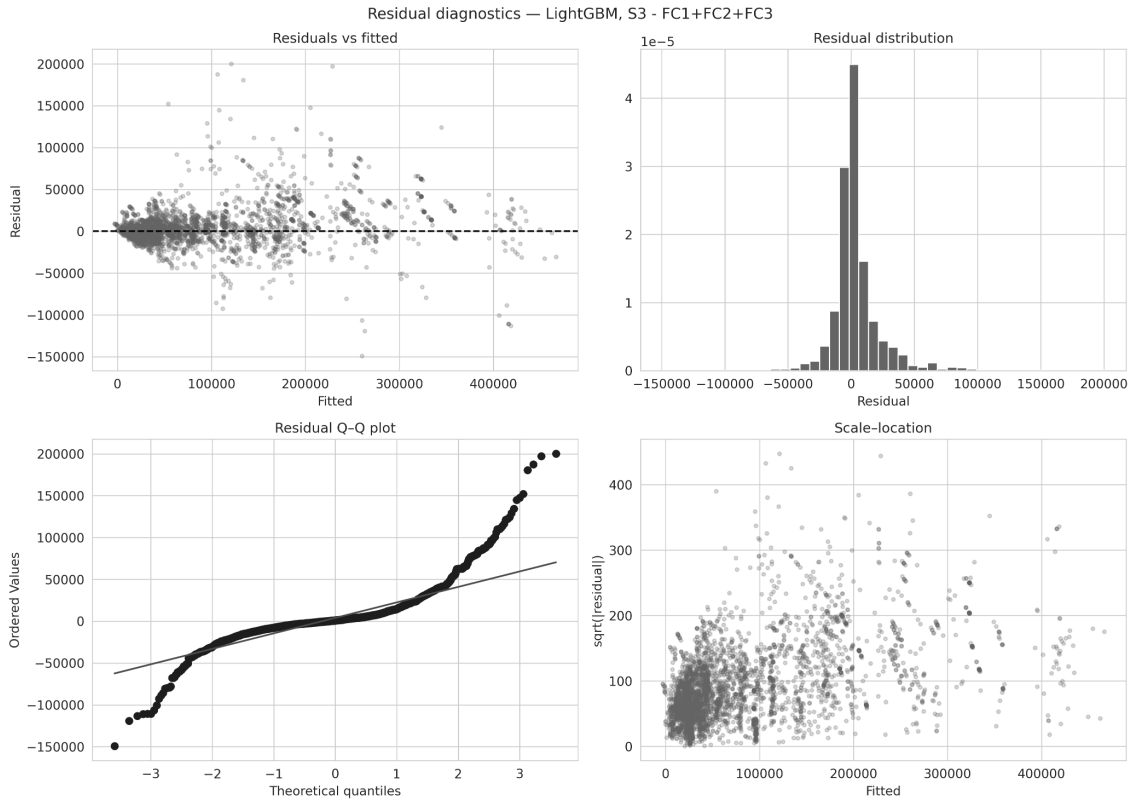


**Figure 9.** Residual diagnostics for S3-LightGBM on the time-based test set

Errors were heterogeneous across countries, years, and crops. Guyana had the highest country-level MAE (47,110 hg/ha), 2013 had the highest annual MAE (14,020 hg/ha), and cassava had the highest crop-level MAE (24,625 hg/ha). Residual spread increased at higher fitted values and the Q-Q plot showed heavy tails. Figure 9 confirms heteroskedasticity and heavy-tailed errors, indicating that average accuracy does not imply constant predictive reliability across the yield range. Table 12 examines whether the relative-error metric changes systematically across observed-yield quartiles.

**Table 12.** MAPE sensitivity across yield quartiles for S3-LightGBM

| Yield quartile | n | Yield range | Mean yield | MAE | MAPE (%) |
|---|---|---|---|---|---|
| Q1 | 1067 | 998-25,401 | 14,208 | 5,533 | 50.59 |
| Q2 | 1031 | 25,449-47,194 | 33,329 | 5,463 | 16.20 |
| Q3 | 1051 | 47,397-118,459 | 75,764 | 10,366 | 14.44 |
| Q4 | 1047 | 118,660-468,991 | 226,614 | 26,383 | 11.67 |

MAPE reached 50.59% in the lowest-yield quartile but only 11.67% in the highest-yield quartile. For the selected time-based model, WMAPE was 13.66% and sMAPE was 20.02%. MAPE is therefore not used as the primary selection metric. Table 12 confirms that low denominators, rather than large absolute errors alone, drive the elevated MAPE in Q1; this justifies emphasizing $R^2$, RMSE, MAE, WMAPE, and sMAPE in model assessment.

### 3.6. Held-Out-Country Validation

The same 19 countries and 761 observations from 2010-2013 were used to compare S1 and S3 after all development-stage decisions were complete. All confidence intervals in Table 13 are based on 2,000 country-cluster bootstrap resamples. Table 13 presents the final, previously unseen-country comparison for S1 and S3 using identical held-out observations.

**Table 13.** Direct S1 versus S3 held-out-country evaluation

| Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE [95% CI] | MAPE (%) |
|---|---|---|---|---|---|
| S1 - Primitive | XGBoost | 0.6382 [0.5494, 0.7362] | 60,080 [45,103, 70,517] | 39,292 [29,295, 47,272] | 80.59 |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Scenario | Model | $R^2$ [95% CI] | RMSE [95% CI] | MAE [95% CI] | MAPE (%) |
|---|---|---|---|---|---|
| S1 - Primitive | LightGBM | 0.6694 [0.5793, 0.7551] | 57,431 [43,243, 67,059] | 37,153 [28,323, 44,163] | 75.32 |
| S3 - FC1+FC2+FC3 | XGBoost | 0.6423 [0.5589, 0.7496] | 59,743 [43,577, 70,441] | 37,954 [27,664, 45,890] | 74.29 |
| S3 - FC1+FC2+FC3 | LightGBM | 0.6169 [0.5271, 0.7431] | 61,823 [44,086, 72,540] | 39,073 [27,498, 47,701] | 79.51 |

S1-LightGBM produced the highest aggregate held-out $R^2$ (0.6694), whereas S3-LightGBM, despite being the temporal winner, declined to 0.6169. For XGBoost, S3 reduced country-level MAE relative to S1 ($d_z$ = 0.629, paired t-test p = 0.0134; Wilcoxon p = 0.0008). For LightGBM, S1 and S3 country-level MAE did not differ significantly ($d_z$ = 0.053, p = 0.819). The configuration favored by temporal validation is therefore not automatically the most transferable to unseen countries. Table 13 demonstrates that temporal superiority does not transfer automatically: S1–LightGBM leads in aggregate held-out $R^2$, while S3 yields a statistically meaningful MAE reduction only for XGBoost. Figure 10 summarizes the Nemenyi average ranks for the eight model–scenario combinations and displays the critical-difference threshold.

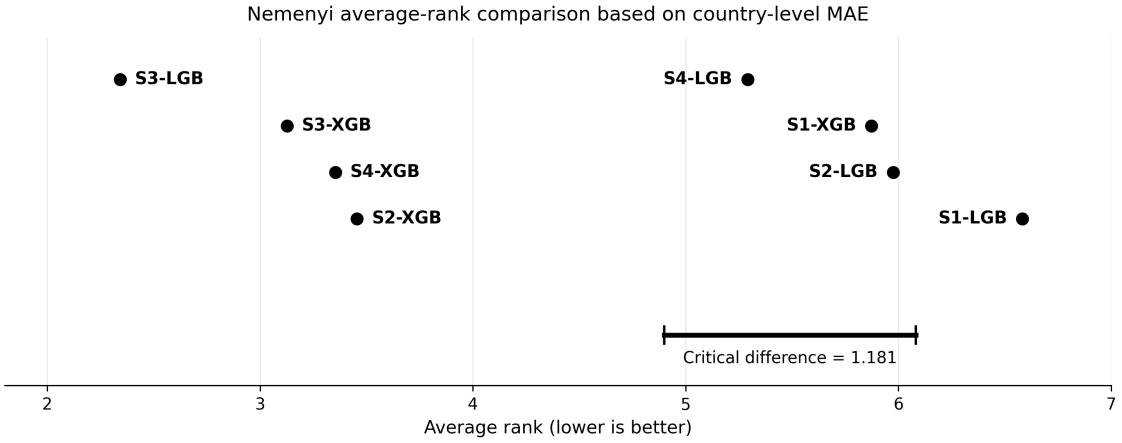


**Figure 10.** Nemenyi critical-difference summary based on country-level MAE

Figure 10 places S3–LightGBM at the best average rank and shows that only rank separations exceeding 1.181 are statistically distinguishable; the diagram therefore complements, rather than replaces, the pairwise p-value matrix.

### 3.7. Discussion

The revised experiment no longer supports a simple conclusion that either primitive variables or the full feature set are always preferable. S3-LightGBM achieved the strongest temporal performance, while isolated modules affected the two algorithms differently. FC-1 reduced both models; FC-3 improved both in isolation; FC-5 gave the largest isolated gain for XGBoost; and FC-2 gave the largest isolated gain for LightGBM. These results indicate that the value of an engineered module depends on the learner and the variables already present. Recent systematic reviews similarly show that the effectiveness of crop-yield predictors depends on the selected input variables, modeling algorithm, evaluation metric, and spatial or temporal context [25], [26].

LightGBM's decline from S3 to S4 is consistent with the added redundancy in FC-4 and FC-5. Several S4 variables are deterministic or near-deterministic transforms of earlier features, including fc1_gdd and fc4_tgsl, rainfall lag and rainfall rolling mean, and the ombrothermic and Lang ratios. LightGBM's leaf-wise strategy can exploit many alternative split candidates on the development data, but those partitions may not transfer to later years. This is a diagnostic explanation rather than a formally isolated causal mechanism.

Random partitioning was optimistic because repeated country-crop series appeared across training and test subsets. Time-based validation removed future-year overlap within development countries, whereas GroupKFold and held-out validation tested geographic transfer. The resulting rankings differed: S3-LightGBM was best over future years within observed countries, S4-LightGBM had the highest individual GroupKFold $R^2$, and S1-LightGBM had the highest held-out-country $R^2$. Accordingly, 'best model' is meaningful only after the intended generalization target has been specified. Recent regional and hybrid-modeling studies also demonstrate that model rankings and predictive robustness can change with geographical coverage, data availability, and the integration of process-based agricultural knowledge [27], [28] .

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The direct S1-S3 held-out comparison also qualifies the earlier inference from GroupKFold. S3 improved XGBoost's country-level error but did not improve LightGBM's held-out performance. Country-grouped cross-validation remains useful, but it cannot guarantee the same ranking on a fixed unseen-country set. The all-zero country indicator for held-out countries further shows that heavy reliance on country one-hot features can limit transfer beyond the development geography.

Crop and country indicators encode characteristic yield distributions and therefore explain a substantial proportion of model behavior. Their importance should not be interpreted as a causal agricultural mechanism. Similarly, $CO_2$ and Year are strongly collinear trend proxies. SHAP values distribute attribution among correlated features according to the fitted model and background distribution; they do not identify independent causal effects [23]. Year, $CO_2$, and deterministic temporal transforms should therefore be interpreted jointly. Similar explainable-AI applications in global crop analysis use feature importance and SHAP to describe fitted-model behavior while cautioning against treating predictive attribution as independent causal evidence [29].

All analyses use one country-level Kaggle dataset described as compiled from FAO and World Bank sources. Shared reporting conventions, national aggregation, and uneven monitoring capacity may influence both development and held-out observations. The 19 held-out countries were selected randomly rather than stratified, and the time-based test covers only four years. The framework has not yet been validated on an independent dataset or on held-out crop types. Future validation could incorporate transfer-learning strategies and spatially explicit yield-potential products, which have recently been used to improve crop-yield estimation across heterogeneous environments and geographical scales [30], [31].

Annual rainfall and mean temperature cannot reproduce daily weather dynamics, soil conditions, cultivar differences, irrigation, fertilizer, or management practices. The GDD, PET, and composite variables are proxies derived from annual aggregates. MAPE is also unstable at low yield levels. Consequently, the present results support methodological benchmarking at country scale, not operational smart-agriculture or farm-level decision support.

## 4. CONCLUSION

This study evaluated a modular agroclimatic feature engineering framework for country-level crop-yield prediction using 28,151 observations from 98 countries. A leakage-controlled protocol reserved 19 countries before feature-statistics fitting, encoding, tuning, scenario selection, and model fitting, and combined time-based, country-grouped, and held-out-country evaluation. The best time-based result was achieved by S3-LightGBM ($R^2$ = 0.9492; RMSE = 21,257 hg/ha), whereas the best held-out-country result was achieved by the simpler S1-LightGBM configuration ($R^2$ = 0.6694; RMSE = 57,431 hg/ha). Direct S1-S3 testing therefore confirmed that temporal and geographic generalization can favor different feature configurations. Per-module ablation further showed that feature effects were model-dependent: FC-1 reduced both models, FC-3 improved both in isolation, FC-5 was strongest for XGBoost, and FC-2 was strongest for LightGBM. Country-cluster bootstrap intervals, country-level effect sizes, and Friedman-Nemenyi analysis supported these configuration-specific differences. The framework is valuable as a country-level benchmarking and diagnostic procedure, but its lower held-out-country accuracy, dependence on one aggregated dataset, proxy agroclimatic variables, correlated temporal features, and sensitivity of MAPE at low yields prevent claims of operational readiness. Future research should use independent crop-yield datasets, daily meteorological observations, soil and management variables, sub-national records, stratified geographic holdouts, and held-out crop validation.

## ACKNOWLEDGMENT

The authors thank the Faculty of Computer Science, Universitas Amikom Yogyakarta, for academic support. The authors also acknowledge the public dataset distributed through Kaggle and its stated FAO and World Bank provenance.

## REFERENCES

[1] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[2] K. S. M. Anbananthen et al., "An intelligent decision support system for crop yield prediction using hybrid machine learning algorithms," *F1000Res.*, vol. 10, p. 1143, Nov. 2021, doi: 10.12688/f1000research.73009.1.

[3] M. Shahhosseini, G. Hu, I. Huber, and S. V. Archontoulis, "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt," *Sci. Rep.*, vol. 11, no. 1, p. 1606, Jan. 2021, doi: 10.1038/s41598-020-80820-1.

[4] M. Hasan et al., "Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation," *Front. Plant Sci.*, vol. 14, Aug. 2023, doi: 10.3389/fpls.2023.1234555.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[6] D. Wang, L. Li, and D. Zhao, "Corporate finance risk prediction based on LightGBM," *Inf. Sci.*, vol. 602, pp. 259–268, Jul. 2022, doi: 10.1016/j.ins.2022.04.058.

[7] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[8] Y. Li et al., "A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 118, p. 103269, Apr. 2023, doi: 10.1016/j.jag.2023.103269.

[9] S. S. Sajid, M. Shahhosseini, I. Huber, G. Hu, and S. V. Archontoulis, "County-scale crop yield prediction by integrating crop simulation with machine learning models," *Front. Plant Sci.*, vol. 13, Nov. 2022, doi: 10.3389/fpls.2022.1000224.

[10] S. M. Abdelouahed, R. Abla, E. Asmae, and A. Abdellah, "Harnessing feature engineering to improve machine learning: A review of different data processing techniques," in 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), IEEE, May 2024, pp. 1–6. doi: 10.1109/ISCV60512.2024.10620105.

[11] A. Morales and F. J. Villalobos, "Using machine learning for crop yield prediction in the past or the future," *Front. Plant Sci.*, vol. 14, Mar. 2023, doi: 10.3389/fpls.2023.1128388.

[12] "Crop Yield Prediction Dataset." Accessed: Aug. 01, 2026. [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[13] T. Iizumi and T. Sakai, "The global dataset of historical yields for major crops 1981–2016," *Sci. Data*, vol. 7, no. 1, p. 97, Mar. 2020, doi: 10.1038/s41597-020-0433-7.

[14] D. Grogan, S. Frolking, D. Wisser, A. Prusevich, and S. Glidden, "Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015," *Sci. Data*, vol. 9, no. 1, p. 15, Jan. 2022, doi: 10.1038/s41597-021-01115-2.

[15] "Trends in CO2 - NOAA Global Monitoring Laboratory." Accessed: Aug. 01, 2026. [Online]. Available: https://gml.noaa.gov/ccgg/trends/data.html

[16] I. Charalampopoulos, F. Droulia, and I. X. Tsiros, "Projecting Bioclimatic Change over the South-Eastern European Agricultural and Natural Areas via Ultrahigh-Resolution Analysis of the de Martonne Index," *Atmosphere*, vol. 14, no. 5, p. 858, May 2023, doi: 10.3390/atmos14050858.

[17] S. Liao et al., "A Modified Shape Model Incorporating Continuous Accumulated Growing Degree Days for Phenology Detection of Early Rice," *Remote Sens.*, vol. 14, no. 21, p. 5337, Oct. 2022, doi: 10.3390/rs14215337.

[18] V. Aschonitis, D. Touloumidis, M.-C. ten Veldhuis, and M. Coenders-Gerrits, "Correcting Thornthwaite potential evapotranspiration using a global grid of local coefficients to support temperature-based estimations of reference evapotranspiration and aridity indices," *Earth Syst. Sci. Data*, vol. 14, no. 1, pp. 163–177, Jan. 2022, doi: 10.5194/essd-14-163-2022.

[19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA: ACM, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[20] J. Carrasco, S. García, M. M. Rueda, S. Das, and F. Herrera, "Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: Practical guidelines and a critical review," *Swarm Evol. Comput.*, vol. 54, p. 100665, May 2020, doi: 10.1016/j.swevo.2020.100665.

[21] S. Herbold, "Autorank: A Python package for automated ranking of classifiers," *J. Open Source Softw.*, vol. 5, no. 48, p. 2173, Apr. 2020, doi: 10.21105/joss.02173.

[22] J. Liu and Y. Xu, "T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, p. 29, Dec. 2022, doi: 10.1007/s44196-022-00083-8.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[23] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[24] R. Salmerón-Gómez, C. B. García-García, and J. García-Pérez, "A Redefined Variance Inflation Factor: Overcoming the Limitations of the Variance Inflation Factor," *Comput. Econ.*, vol. 65, no. 1, pp. 337–363, Jan. 2025, doi: 10.1007/s10614-024-10575-8.

[25] S. M. Shawon, F. B. Ema, A. K. Mahi, F. L. Niha, and H. T. Zubair, "Crop yield prediction using machine learning: An extensive and systematic literature review," *Smart Agric. Technol.*, vol. 10, p. 100718, Mar. 2025, doi: 10.1016/j.atech.2024.100718.

[26] Md. A. Jabed and M. A. Azmi Murad, "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability," *Heliyon*, vol. 10, no. 24, p. e40836, Dec. 2024, doi: 10.1016/j.heliyon.2024.e40836.

[27] M. A. Razavi et al., "Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data," *Artif. Intell. Agric.*, vol. 14, pp. 99–114, Dec. 2024, doi: 10.1016/j.aiia.2024.11.005.

[28] M. von Bloh, D. Lobell, and S. Asseng, "Knowledge informed hybrid machine learning in agricultural yield prediction," *Comput. Electron. Agric.*, vol. 227, p. 109606, Dec. 2024, doi: 10.1016/j.compag.2024.109606.

[29] Y. Li et al., "Research on Factors Affecting Global Grain Legume Yield Based on Explainable Artificial Intelligence," *Agriculture*, vol. 14, no. 3, p. 438, Mar. 2024, doi: 10.3390/agriculture14030438.

[30] S. N. Khan, D. Li, and M. Maimaitijiang, "Using gross primary production data and deep transfer learning for crop yield prediction in the US Corn Belt," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 131, p. 103965, Jul. 2024, doi: 10.1016/j.jag.2024.103965.

[31] F. Aramburu-Merlos, M. P. van Loon, M. K. van Ittersum, and P. Grassini, "High-resolution global maps of yield potential with local relevance for targeted crop production improvement," *Nat. Food*, vol. 5, no. 8, pp. 667–672, Jul. 2024, doi: 10.1038/s43016-024-01029-3.