Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 2, No. 1, [June 2023], pp. [1–10]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Tree-Structured Parzen Estimator–Optimized XGBoost with Spatial Cross-Validation for Improved Relative Soil Fertility Classification

Hilmi Aziz Bukhori [1*], Lidya Wati [2], Tutut Herawan [3], Sankar Kumar Roy [4]

[1] *Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia*
[2] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[3] *Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia.*
[4] *Department of Applied Mathematics, Vidyasagar University, Midanpore, India*

[1] *hilmibukhori@ub.ac.id;* [2] *lidyawati@studentmail.unimap.edu.my;* [3] *tutut@um.edu.my;* [4] *sankroy2006@gmail.com*
** Corresponding author: hbukhori@bcamath.org.*

Abstract

This study examines XGBoost configuration and deterministic feature-representation search for relative soil fertility classification. A 20,000-row sample is drawn from a 135,143-row local dataset. The original Fertility_Level is constant; three relative labels are instead defined by outer-training Fertility tertiles, with thresholds 431 and 485. The coordinate-block split contains 15,747 training and 4,253 test observations. The code specifies three-fold group cross-validation and 32 TPE trials, including nine random startup trials. It searches tree settings and raw or augmented representations, with auxiliary features standardized using the complete outer-training data. Stored test accuracy rises from 90.03% to 94.87%, and macro-F1 from 86.44% to 92.49%. Class-1 recall increases from 74.81% to 84.73%. The archived ablation summary indicates a substantial representation-associated difference, while tuning raw inputs alone does not improve the fixed baseline. The winning settings, feature type and trial history are absent, and the historical test was previously inspected. The results therefore support a retrospective improvement for the recorded selected model, without establishing a separate TPE effect, agronomic label validity or independent geographic generalization.

Keywords: relative soil fertility; XGBoost; spatial cross-validation; model selection; macro-F1

1. Introduction

The growing demand for spatial soil information has encouraged the use of machine learning to support fertility assessment and soil management. Recent reviews describe a broad range of learning approaches for this purpose [1], [2], while soil science-informed modeling emphasizes that a predictive relationship must remain consistent with the meaning and limitations of the target [3]. Soil-fertility-index prediction and real-time fertility analysis illustrate different application settings [4], [5]. These contributions motivate investigating relative fertility classification, but they do not make a class formed from a numerical score equivalent to an independently validated agronomic diagnosis. The present study therefore distinguishes the supplied fertility measurements from the three relative labels constructed for the experiment. The choice of predictors and evaluation design is central to interpreting a reported improvement. Studies of UAV hyperspectral nutrient monitoring, spatial-neighbor information and parent-material covariates examine different sources of predictive information [6], [7], [8]. Greedy selection, recursive variable selection and broader feature-selection research show why the input representation should be stated explicitly [9], [10], [11]. General feature-engineering and scaling studies likewise treat preprocessing as part of the learning procedure [12], [13]. These findings motivate documenting the available predictors and every transformation rather than attributing a higher score to an unspecified feature intervention. In the current

experiment, Latitude, Longitude, NDVI, Soil_Moisture and Elevation are the only five original predictors; Fertility and Fertility_Level are excluded from the model input.

Spatial dependence and model selection require separate attention. Soil-profile leakage, digital-soil-mapping partition experiments and extrapolation studies show that performance depends on the validation unit and prediction setting [14], [15], [16]. Spatial validation and clustered-sample assessment address related concerns [17], [18], [19]. Hyperparameter-optimization and automated-learning research further requires an explicit search domain, objective and budget [20], [21], [22]. The present comparison uses one coordinate-block holdout and training-only group cross-validation. Its purpose is to assess the recorded model configurations under that partition, with macro-F1 as the selection endpoint. This scope does not establish transfer to a new region, calibration quality or a fertilizer recommendation. The present study examines joint selection of XGBoost settings and a deterministic feature representation under three-fold group cross-validation. It compares a fixed raw-feature baseline with the stored Optimized evaluation and reports the available four-row ablation summary. The search is configured for 32 trials with nine random startup trials. The archived evaluation does not embed the winning parameter vector or feature type, so the contribution is a retrospective configuration comparison with explicit limits on component attribution.

2. Method

2.1. Related Work

Recent soil-fertility reviews synthesize models for assessing fertility and supporting management decisions [1], [2]. Soil science-informed machine learning places the scientific target and domain knowledge within that modeling process [3]. The urban Loess Plateau study predicts a soil fertility index using tree-based models [4], while a real-time analysis study considers fertility analysis and crop-related outputs [5]. These studies concern different targets, measurements and application settings. Their performance estimates cannot validate the relative classes used here, which are defined by training-set quantiles of a supplied Fertility score. Environmental representation has been examined through UAV hyperspectral nutrient monitoring [6], explicit soil spatial-neighbor information [7], and the addition of parent-material information in Bogor [8]. These interventions supply different information to a model. The current five-variable input does not contain hyperspectral bands, neighboring measured soil responses or a parent-material map. Consequently, their reported benefits provide context for representation design rather than evidence that the current predictors recover the same soil processes. Feature-selection research in digital soil mapping includes modified greedy selection for parsimonious models and recursive variable selection for multiple soil properties [9], [10]. A recent high-dimensional study evaluates interactions between metaheuristic selection, model architecture and prediction target [23]. General surveys distinguish selection of an existing subset from construction of new predictors [11], [12]. This distinction matters when a method title mentions feature selection: an evaluated search procedure does not by itself establish that its selected model uses fewer inputs or that the selected subset is stable. Ensemble-based micronutrient mapping provides an application of combined learners to regional soil information [24]. A comparative soil-pH study evaluates tree models together with residual kriging, and Bayesian–machine-learning synergy for soil organic matter combines spatial modeling with a machine-learning predictor [25], [26] . Multi-feature WMTR-CatBoost modeling addresses soil organic carbon and nitrogen [27]. These are useful examples of changing the learner or integrating complementary information. Their continuous targets, covariate sets and spatial treatments differ from the present three-class experiment; they do not imply that a particular ensemble or boosting implementation must win here. Hyperparameter optimization treats the search algorithm, domain, objective and evaluation budget as elements of one procedure [20]. Automated-learning reviews place these choices alongside preprocessing and learner selection, and empirical tree-tuning research demonstrates the importance of stating the effective tree configuration [21], [22]. A bounded search should therefore be distinguished from a fully documented winning run. The number of configured trials alone does not recover the trial history, selected parameters or variability of repeated optimization.

Tao and colleagues use TPE to select a representation-learning hyperparameter for imbalanced tabular data [28]. This supplies a recent journal example of TPE within model development, but it is not an XGBoost soil benchmark. Here, TPE is configured to search both tree settings and five eligible representations. The standardized auxiliary features are computed using the full outer-training mean and standard deviation before the inner folds. Scaling research treats this choice as part of the input procedure [13]; it does not establish that the transformation improves this experiment or that inner-fold preprocessing is fully isolated. Leakage research describes how preprocessing, model selection and repeated inspection can compromise the interpretation of an assessment [29], [30], [31]. REFORMS provides recommendations for reporting machine-learning-based science [32]. These contributions motivate recording the exact development boundary and retaining the data, search settings and final model

information. In the current study, the relative-label thresholds are obtained from the complete outer training set before inner cross-validation; they are not re-estimated separately in each inner fold. Thus, the reported CV objective is conditional on that outer-training target definition. Method-comparison research explains why benchmarks and comparators affect the conclusions drawn from an experiment [33], [34]. Cross-validation theory distinguishes a development estimate from the behavior of a final fitted procedure [35]. For this study, the relevant contrast is between the explicitly recorded configurations on the same held-out rows. A single split and one search seed do not establish a general ranking of algorithms, and differences between differently budgeted procedures do not isolate an optimizer effect. The soil-profile leakage study specifically considers dependent observations within profiles , whereas partition-randomness research examines the effect of resampling choices in digital soil mapping [14], [15]. Spatial+ and clustered-sample assessment address geographic dependence under their respective validation designs [17], [18]. Block-design research further shows that the chosen block structure affects what a spatial test means [19]. Here, coordinate-based group identifiers are kept disjoint across the outer split and across inner folds. No distance buffer is implemented, so disjoint group codes are not interpreted as a guarantee of independence at every spatial scale. Multiclass metric guidance distinguishes overall accuracy from class-balanced and support-weighted summaries [36], [37]. Learning-curve research considers changes in predictive behavior with fitting-set size [38] . These studies motivate reporting macro-F1, balanced accuracy, per-class recall and confusion counts, together with the recorded internal learning curves. The curves reuse development observations after configuration selection, and therefore serve as diagnostics rather than independent confirmation of the final result. Uncertainty validation in digital soil mapping concerns the reliability of probabilistic predictions, classifier-calibration research similarly distinguishes probability reliability from hard-label correctness [39], [40]. The review of explainable spectral soil-fertility modeling examines another dimension of model interpretation [41]. Deployment studies document additional constraints beyond an accuracy table [42]. These contributions define questions requiring separate evidence. No calibration experiment, explanation analysis, operational timing benchmark or field-level decision evaluation is inferred from the stored confusion matrices.

2.2. Dataset

The experiment uses the local Sakri_Soil_Fertility_Data.csv file, which contains 135,143 rows and seven columns. A deterministic pandas sample of 20,000 rows is drawn with random_state=42 and its row index is reset. The five numerical predictors are Latitude, Longitude, NDVI, Soil_Moisture and Elevation. Fertility is used only to form the response, and the supplied Fertility_Level column is constant at High for every source row. Therefore, the study does not use the originally proposed four-category target. Table 1 records the actual dataset and partition.

Table 1. Verified local data and experimental partition

Attribute	Observed specification
Source file	135,143 rows × 7 columns; no missing values in the seven required fields
Experiment sample	20,000 rows; pandas sample random_state=42
Inputs / excluded columns	Five numerical inputs; Fertility and Fertility_Level excluded from predictors
Relative target	Class 0: $F < 431$; class 1: $431 \leq F < 485$; class 2: $F \geq 485$
Training partition	15,747 rows; 78 encoded blocks; class counts 5,184 / 5,235 / 5,328
Test partition	4,253 rows; 22 encoded blocks; class counts 2,402 / 786 / 1,065

The original dataset label is uninformative for classification because every row is High. The notebook instead calculates the one-third and two-thirds quantiles of Fertility using only the outer training rows, obtaining 431 and 485. The same thresholds label the held-out rows, with values equal to a threshold placed in the higher class. The terms class 0, class 1 and class 2 are used throughout to emphasize their relative meaning. The source does not establish an agronomic unit, field-validated fertility threshold or 30 m observation resolution. The notebook names https://data.mendeley.com/datasets/h73wn2nw56/1 as a candidate source while explicitly stating that identity with the local Sakri file is unverified. This analysis therefore identifies the local file and its actual row counts rather than attributing the experiment to that public record. The source-file SHA-256 is b212a3791007499648ef9c1faebe7018d15253560c0b993a8ac46eb669a54ed7.

Table 2. Implemented development-to-test flow and its limits

Stage	Implemented handling
Sample and group	Sample 20,000 rows; form coordinate-grid codes using 0.01-degree latitude/longitude bins
Outer partition	GroupShuffleSplit: n_splits=1, test_size=0.22, random_state=42; test_size is a group fraction
Define labels	Training Fertility quantiles only; fixed thresholds reused for inner CV and outer test
Inner validation	GroupKFold with three folds; validation sizes 5,253, 5,247 and 5,247

Stage	Implemented handling
Final comparison	Refit selected configurations on the outer training rows and retain the stored test outputs

The outer train and test group identifiers are disjoint, as are the training and validation groups in each inner fold. The 0.22 split setting refers to groups, so it yields 4,253 rows rather than exactly 22% of 20,000 rows. The grouping has no explicit buffer or distance-based exclusion zone. Group codes, relative labels and a single outer split define the tested task; they do not establish an independently sampled new geographic region.

2.3. Research Framework

Figure 1 summarizes the common data boundary and the recorded comparison arms. Every arm uses the same five original covariates, training-defined target and outer test membership. The selected objective is based on the training group folds; the outer test metrics are retained as the final stored comparison.

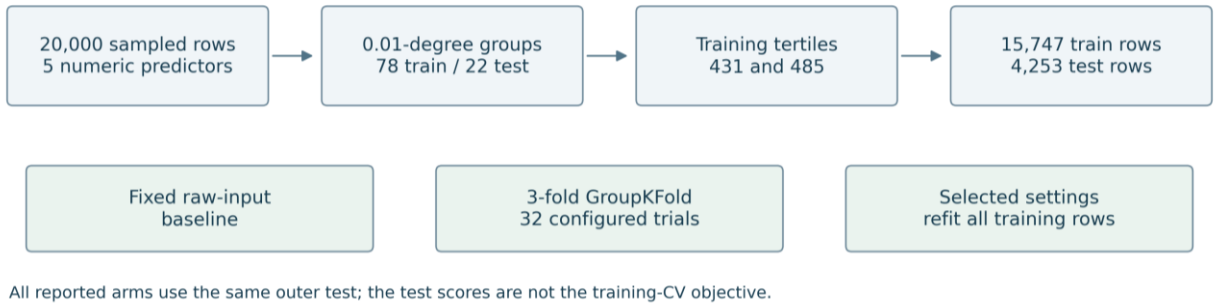

Figure 1. Shared coordinate-block data boundary and comparison arms

XGBClassifier: 200 boosting rounds, max_depth=6, learning_rate=0.1, tree_method=hist, objective=multi:softprob, eval_metric=mlogloss, n_jobs=2 and random_state=42.

Table 3. Recorded comparison and controlled factors

Component	Study-specific implementation
Fixed comparator	XGBClassifier: 200 boosting rounds, max_depth=6, learning_rate=0.1, tree_method=hist, objective=multi:softprob, eval_metric=mlogloss, n_jobs=2 and random_state=42.
Selection arm	Joint search over XGBoost hyperparameters and feature_type ∈ {raw, coords, soil, full, geo}; 32 configured trials, nine random startup trials, seed 42.
Shared controls	Same sampled rows, coordinate groups, relative-label thresholds and outer test cases
Reported endpoints	Accuracy, macro-F1, weighted F1, balanced accuracy and class-wise confusion counts
Interpretation boundary	One recorded partition and seed; no independent agronomic validation or repeated-search uncertainty

Figure 2 details the implemented representation, selection and refitting stages. The equations in the Proposed Method make the target, grouping, prediction rule and search objective explicit. They describe the supplied code and do not introduce a new experiment.

2.4. Proposed Method

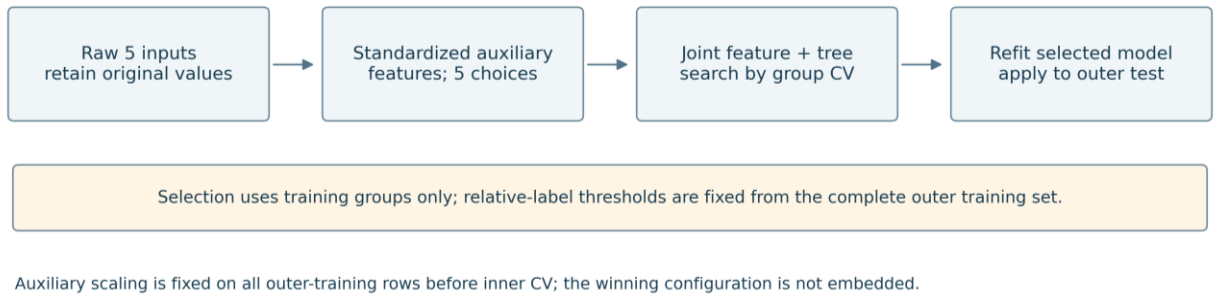

Figure 2. Implemented method and selection boundary

Table 4. Search domain and evaluation controls

Component	Configured domain or control
Representation	raw (5), coords (11), soil (14), full (20), geo (14)
Boosting rounds	140, 180, …, 500; the coded upper bound 520 is not reached with step 40
Tree structure	max_depth: 3–12; min_child_weight: integer 1–12
Shrinkage / sampling	learning_rate: 0.025–0.25 (log); subsample and colsample_bytree: 0.70–1.00
Regularization	gamma: 0–2; reg_lambda: 0.01–30 (log); reg_alpha: 0.0001–5 (log)
Objective / budget	Mean three-fold group-CV macro-F1; 32 requested trials; nine startup trials; seed 42

Let $\mathcal{T}$ and $\mathcal{E}$ denote the fixed outer training and test indices. For numerical Fertility score F_i, the empirical training quantiles in (1) define the task. The indicator $\mathbf{1}[P]$ equals one when statement P is true. Equality with a threshold enters the higher class, matching searchsorted with side set to right. These labels express position relative to the training distribution, not validated agronomic thresholds.

$$q_1 = Q_{1/3}(F_{\mathcal{T}}) = 431, \quad q_2 = Q_{2/3}(F_{\mathcal{T}}) = 485, \quad y_i = \mathbf{1}[F_i \geq q_1] + \mathbf{1}[F_i \geq q_2]. \tag{1}$$

For latitude a_i and longitude b_i, (2) gives the coordinate-grid code used for grouping. The minima are computed on the 20,000-row sample, before splitting; the grid width is 0.01 degrees. GroupShuffleSplit assigns complete codes to either outer partition, and GroupKFold uses the training codes for the three inner folds. No geographic distance buffer is applied.

$$u_i = \left\lfloor \frac{a_i - \min_j a_j}{0.01} \right\rfloor, \quad v_i = \left\lfloor \frac{b_i - \min_j b_j}{0.01} \right\rfloor, \quad g_i = 1000u_i + v_i. \tag{2}$$

Write the original predictor vector as $x = (a, b, n, m, e)^T$, containing Latitude, Longitude, NDVI, Soil_Moisture and Elevation. For auxiliary-feature construction, (3) standardizes each component using the complete outer training set. Here $N = |\mathcal{T}|$ and the standard deviation uses the population denominator, as in NumPy std with its default ddof=0. The original five inputs are retained on their original scales. This transformation is fixed before inner CV, so validation-fold covariates contribute to its training mean and scale.

$$\mu_j = \frac{1}{N}\sum_{i\in\mathcal{T}} x_{ij}, \quad s_j = \sqrt{\frac{1}{N}\sum_{i\in\mathcal{T}} \left(x_{ij} - \mu_j\right)^2}, \quad z_j = \frac{x_j - \mu_j}{s_j}. \tag{3}$$

Let $z = (A, B, N_d, M, E)^T$ denote the five standardized components, with N_d denoting standardized NDVI rather than sample count. Equation (4) defines all 15 auxiliary features in their code order. These are algebraic coordinate and environmental interactions; the code does not add neighboring soil observations, new sensor bands or new field measurements.

$$\begin{aligned} \psi_{(1:8)}(x) &= [A + B,\ A - B,\ AB,\ A^2,\ B^2,\ N_d M,\ N_d E,\ ME]^T, \\ \psi_{(9:15)}(x) &= \left[N_d^2,\ M^2,\ E^2,\ \sqrt{A^2 + B^2},\ |N_d - M|,\ N_d + M,\ N_d - M\right]^T. \end{aligned} \tag{4}$$

Using one-based indices of ψ, the candidate representations in (5) match the five feature_type choices. Concatenation preserves all five raw values; parentheses list selected auxiliary indices. The arrays are cast to float32. The chosen representation is not embedded in the archived evaluation.

$$\begin{aligned} \phi_{\text{raw}}(x) &= x \in \mathbb{R}^5, \\ \phi_{\text{coords}}(x) &= \left[x^T, \psi_{(1,2,3,4,5,12)}^T\right]^T \in \mathbb{R}^{11}, \\ \phi_{\text{soil}}(x) &= \left[x^T, \psi_{(6,7,8,9,10,11,13,14,15)}^T\right]^T \in \mathbb{R}^{14}, \\ \phi_{\text{full}}(x) &= [x^T, \psi^T]^T \in \mathbb{R}^{20}, \\ \phi_{\text{geo}}(x) &= \left[x^T, \psi_{(1,2,3,4,5,12,13,14,15)}^T\right]^T \in \mathbb{R}^{14}. \end{aligned} \tag{5}$$

For class $k \in \{0,1,2\}$, let h_{tk} be the class-score tree at round t, b_k the initial score and η the learning rate. Equation (6) gives the additive score and softmax decision rule, with B boosting rounds and candidate representation ϕ_v. The resulting probability estimates are used for classification and are not externally calibrated in this experiment.

$$\begin{aligned} s_k(x) &= b_k + \eta \sum_{t=1}^{B} h_{tk}\left(\phi_v(x)\right), \\ p_k(x) &= \frac{\exp(s_k(x))}{\sum_{c=0}^{2}\exp(s_c(x))}, \qquad \hat{y}(x) = \underset{k\in\{0,1,2\}}{\arg\max}\, p_k(x). \end{aligned} \tag{6}$$

The multiclass fitting loss is negative log likelihood with tree regularization. Equation (7) states the objective form: $\mathcal{A}$ is the candidate fitting fold, L_{tk} the number of leaves, and $w_{tk\ell}$ a leaf score. The nonnegative parameters γ, λ and α correspond to gamma, reg_lambda and reg_alpha. Histogram-based tree fitting updates this model stagewise, with depth, minimum child Hessian weight and sampling constraints; the equation does not assert a globally optimal tree ensemble.

$$\begin{aligned} \mathcal{L} &= -\sum_{i\in\mathcal{A}} \log p_{y_i}(x_i) + \sum_{t=1}^{B}\sum_{k=0}^{2} \Omega\left(h_{tk}\right), \\ \Omega(h_{tk}) &= \gamma L_{tk} + \frac{\lambda}{2}\sum_{\ell=1}^{L_{tk}} w_{tk\ell}^2 + \alpha \sum_{\ell=1}^{L_{tk}} |w_{tk\ell}| \,. \end{aligned} \tag{7}$$

For any evaluation fold, let TP_k, FP_k and FN_k be the class-specific true-positive, false-positive and false-negative counts. Equation (8) defines the unweighted three-class macro-F1 used for model selection. Accuracy and support-weighted F1 are additional reporting endpoints, while balanced accuracy is the mean of the three recalls.

$$\mathrm{MacroF1}(y, \hat{y}) = \frac{1}{3}\sum_{k=0}^{2} \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}. \tag{8}$$

Let $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ be the fixed group-disjoint validation folds within $\mathcal{T}$, and let $\mathcal{T}_{-f}$ contain the rows in the other two folds. For candidate settings θ and representation v, $\hat{y}^{(-f;v,\theta)}$ denotes predictions from a model fitted on $\mathcal{T}_{-f}$. The objective in (9) averages fold-specific macro-F1 values rather than scoring the held-out outer test. Feature standardization and label thresholds remain fixed from the complete outer training set as described above.

$$J(v, \theta) = \frac{1}{3}\sum_{f=1}^{3} \mathrm{MacroF1}\left(y_{\mathcal{I}_f}, \hat{y}_{\mathcal{I}_f}^{(-f;v,\theta)}\right). \tag{9}$$

If $\mathcal{H}_{\mathrm{done}}$ denotes configurations with completed evaluations, (10) specifies selection among them. After selection, the chosen representation and model settings are fixed for refitting on all outer-training rows. The archived test scores do not identify the maximizing pair.

$$(v^*, \theta^*) = \underset{(v,\theta)\in\mathcal{H}_{\mathrm{done}}}{\arg\max}\; J(v, \theta). \tag{10}$$

After the nine random startup trials, the configured TPE sampler proposes candidates using estimated densities of parameters associated with better and remaining objective values. Writing these densities as $\ell(\theta, v)$ and $g(\theta, v)$, the density-ratio criterion in (11) describes its preference among sampled proposals. It is a proposal rule, not proof of a global maximum or of superiority over random search.

$$(v_{\mathrm{next}}, \theta_{\mathrm{next}}) \approx \underset{(v,\theta)\in\mathcal{C}}{\arg\max}\; \frac{\ell(\theta,v)}{g(\theta,v)}. \tag{11}$$

The attached notebook refers to xgb_best.json and xgb_trials.csv but does not embed these files or their contents. The winning parameter vector, feature type, fold scores and completed-trial history remain unavailable. The selection cell also references an undefined SEED, and a later display references an undefined n; these source defects prevent treating the attached notebook as a complete executable reproduction of the stored result.

The notebook warns that the historical test had already been inspected. The stored scores are therefore treated as retrospective evidence, rather than an untouched confirmatory evaluation. Bootstrap code or export statements do not supply an uncertainty interval when no corresponding numerical output is embedded.

3. Results and Discussion

3.1. Overall Classification Performance

The following results are taken from the embedded notebook outputs. Aggregate and class-level metrics are cross-checked against the archived confusion-matrix counts. No estimator is retrained to replace or improve the reported outcome. All models in Table 5 are evaluated on the same 4,253 outer-test observations.

Table 5. Recorded outer-test classification performance

Model	Accuracy (%)	Macro-F1 (%)	Weighted F1 (%)	Balanced acc. (%)
Baseline	90.03	86.44	90.12	86.90
Optimized	94.87	92.49	94.84	92.28

Optimized correctly classifies 4,035 of 4,253 observations, compared with 3,829 for Baseline. Macro-F1 increases by 6.05 percentage points and accuracy by 4.84 points. Balanced accuracy rises from 86.90% to 92.28%. These values describe the two stored configurations on one shared partition; they are not repeated-seed estimates or evidence that TPE alone caused the gain.

3.2. Class-Level Errors

Table 6 reports class recall and F1 on the 0–1 scale. Test supports are 2,402, 786 and 1,065 for relative classes 0, 1 and 2. Class 1 is the smallest test class and remains a major source of confusion. These labels describe training-defined Fertility intervals, not field-validated low, medium or high fertility.

Table 6. Per-class test metrics reconstructed from recorded counts

Model	Class	Support	Precision	Recall	F1
Baseline	0	2402	0.9659	0.9434	0.9545
Baseline	1	786	0.7224	0.7481	0.7350
Baseline	2	1065	0.8920	0.9155	0.9036
Optimized	0	2402	0.9826	0.9896	0.9861
Optimized	1	786	0.8717	0.8473	0.8594
Optimized	2	1065	0.9271	0.9315	0.9293

Class 0 has 2,377 correct predictions in Optimized versus 2,266 in Baseline, and its assignments to class 1 fall from 136 to 25. Correct class-1 predictions increase from 588 to 666, while class-1 assignments to classes 0 and 2 fall from 80 to 42 and from 118 to 78. Class-2 correct predictions increase from 975 to 992. Thus, all three recalls improve, with the largest change in class 1, although that class remains the weakest by F1.

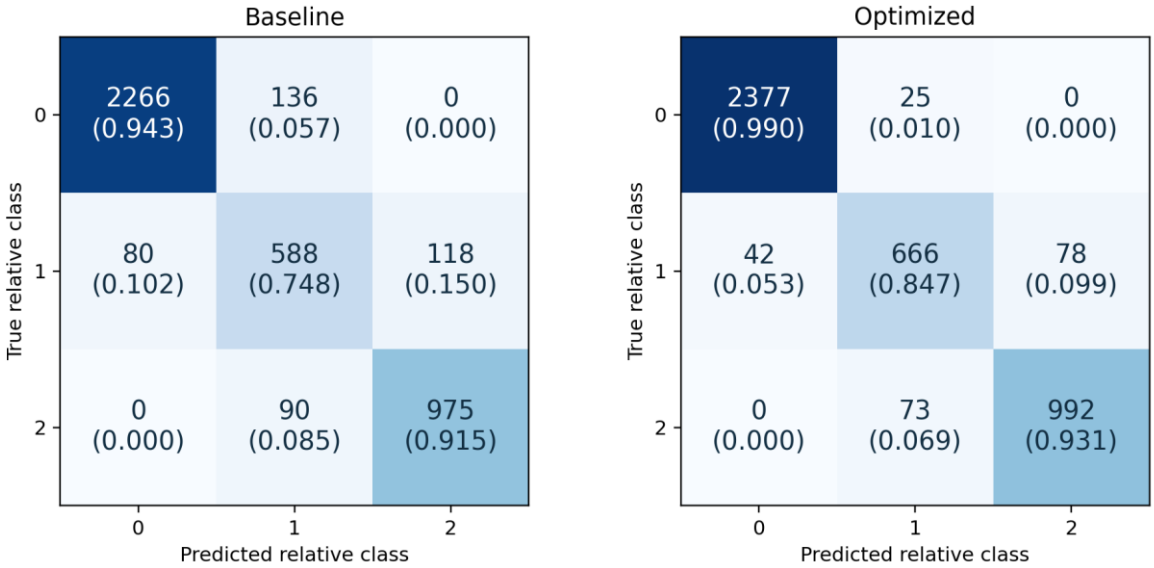

Figure 3. Test confusion matrices; entries show counts and row-normalized proportions

3.3. Internal Learning Curves

Figure 4 reproduces the recorded internal training and validation curves. Fitting subsets contain 4,144, 7,696 and 11,840 observations; the internal validation partition contains 3,907 rows. The split is made inside the outer training data using GroupShuffleSplit, with test_size=0.24 and random_state=54. Smaller fitting subsets are stratified row samples from the internal fitting pool. Selected settings remain fixed during curve refits.

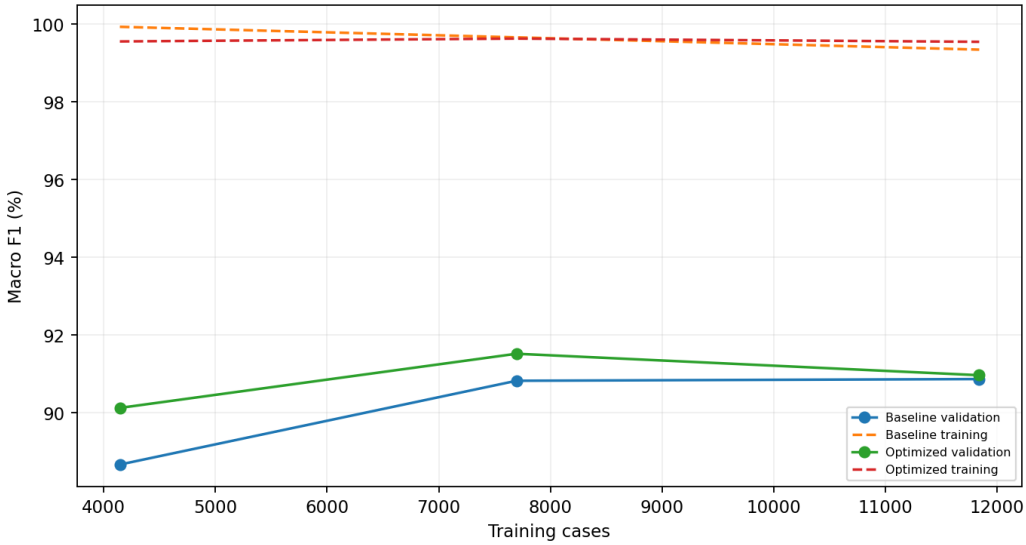

Figure 4. Archived training and internal-validation macro-F1 learning curves

The recorded validation curves increase from the smallest to the intermediate fitting size and then fall slightly at the largest size. Optimized remains above Baseline at all three displayed sizes. Training macro-F1 stays near 100%, leaving a visible training–validation gap. This diagnostic does not demonstrate monotonic improvement with additional data or quantify uncertainty across new spatial splits.

3.4. Discussion and Limitations

The ablation summary associates most of the gain with the engineered-input arm: Baseline_Engineered already reaches 92.30% macro-F1, compared with 92.49% for TPE_Engineered. TPE_Raw is below Baseline_Raw. However, the generating ablation code, full settings and selected feature definitions for these rows are not archived, so the table cannot establish a clean causal decomposition or a general advantage of the optimizer.

Table 7. Stored ablation summary; full generating configurations are not embedded

Stored arm	Accuracy (%)	Macro-F1 (%)
Baseline_Raw	90.03	86.44
Baseline_Engineered	94.78	92.30
TPE_Raw	89.77	85.97
TPE_Engineered	94.87	92.49

The target is a discretization of a supplied numerical Fertility score, and its agronomic validity is not established. The provenance match to the public dataset link remains unverified. The evaluation covers one seed and one coordinate-block split without an explicit spatial buffer; no repeated-seed confidence interval, calibrated probability assessment, feature-importance stability or hardware-specific deployment benchmark is reported. Learning curves reuse development observations after configuration selection. The auxiliary-feature mean and scale are estimated on all outer-training rows before inner CV, and the target thresholds also use all outer-training outcomes. Consequently, the CV score is conditional on these fixed outer-training definitions, rather than a fully nested estimate of the complete preprocessing-and-selection procedure. Prior historical-test inspection further limits confirmation. These limits require a bounded interpretation of the observed configuration differences.

4. Conclusion

The stored XGBoost comparison improves held-out macro-F1 from 86.44% to 92.49% and accuracy from 90.03% to 94.87% under one coordinate-block partition. All three class recalls improve, and the available ablation summary suggests that representation changes account for a substantial part of the recorded difference. The winning representation and hyperparameters are not embedded, and tuning raw inputs alone does not outperform the raw baseline. Full outer-training preprocessing before inner CV, incomplete execution records and prior historical-test inspection limit the interpretation. The findings support a retrospective result for relative labels; complete search records, fold-contained preprocessing, repeated seeds and an independently collected geographic test are needed for stronger claims.

References

[1] M. Sujatha and C. D. Jaidhar, "Machine learning-based approaches to enhance the soil fertility-A review," *Expert Syst. Appl.*, vol. 240, p. 122557, 2024, doi: 10.1016/j.eswa.2023.122557.

[2] S. Muthulakshmi, M. R. Backiyavathy, M. Gopalakrishnan, R. Kalpana, C. Thangamani, and S. I. Pavithra, "Insights of Machine Learning Approach for Soil Fertility Assessment and Management Strategy," *Commun. Soil Sci. Plant Anal.*, vol. 56, no. 3, pp. 436–457, 2025, doi: 10.1080/00103624.2024.2416920.

[3] B. Minasny *et al.*, "Soil Science-Informed Machine Learning," *Geoderma*, vol. 452, p. 117094, 2024, doi: 10.1016/j.geoderma.2024.117094.

[4] X. Shen *et al.*, "Assessment and Prediction of Soil Fertility in Urban Areas of the Loess Plateau Based on Machine Learning Methods," *Clean (Weinh).*, vol. 53, no. 9, p. e70039, 2025, doi: 10.1002/clen.70039.

[5] K. Gunasekaran, K. A., and P. Sreevardhan, "Real-time soil fertility analysis, crop prediction, and insights using machine learning and deep learning algorithms," *Frontiers in Soil Science*, vol. 5, p. 1652058, 2025, doi: 10.3389/fsoil.2025.1652058.

[6] K. Liu *et al.*, "Monitoring soil nutrients using machine learning based on UAV hyperspectral remote sensing," *Int. J. Remote Sens.*, vol. 45, no. 14, pp. 4897–4921, 2024, doi: 10.1080/01431161.2024.2371618.

[7] Z. Chen, Z. Wang, X. Wang, Z. Shi, and S. Chen, "Including soil spatial neighbor information for digital soil mapping," *Geoderma*, vol. 451, p. 117072, 2024, doi: 10.1016/j.geoderma.2024.117072.

[8] D. Cahyana, Y. Sulaeman, B. Barus, Darmawan, and B. Mulyanto, "Improving digital soil mapping in Bogor, Indonesia using parent material information," *Geoderma Regional*, vol. 33, p. e00627, 2023, doi: 10.1016/j.geodrs.2023.e00627.

[9] X. Zhang *et al.*, "Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping," *Geoderma*, vol. 432, p. 116383, 2023, doi: 10.1016/j.geoderma.2023.116383.

[10] J. Huang *et al.*, "Mapping Soil Properties in the Haihun River Sub-Watershed, Yangtze River Basin, China, by Integrating Machine Learning and Variable Selection," *Sensors*, vol. 24, no. 12, p. 3784, 2024, doi: 10.3390/s24123784.

[11] M. R. Islam, A. A. Lima, S. C. Das, M. F. Mridha, A. R. Prodeep, and Y. Watanobe, "A Comprehensive Survey on the Process, Methods, Evaluation, and Challenges of Feature Selection," *IEEE Access*, vol. 10, pp. 99595–99632, 2022, doi: 10.1109/access.2022.3205618.

[12] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, p. 257, 2025, doi: 10.3390/ai6100257.

[13] L. B. V de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, p. 109924, 2023, doi: 10.1016/j.asoc.2022.109924.

[14] K. John, D. D. Saurette, and B. Heung, "The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping," *Geoderma*, vol. 455, p. 117223, 2025, doi: 10.1016/j.geoderma.2025.117223.

[15] D. Radočaj, M. Jurišić, I. Plaščak, and L. Galić, "Randomness in Data Partitioning and Its Impact on Digital Soil Mapping Accuracy: A Comparison of Cross-Validation and Split-Sample Approaches," *Agronomy*, vol. 15, no. 11, p. 2495, 2025, doi: 10.3390/agronomy15112495.

[16] F. Hateffard, L. Steinbuch, and G. B. M. Heuvelink, "Evaluating the extrapolation potential of random forest digital soil mapping," *Geoderma*, vol. 441, p. 116740, 2024, doi: 10.1016/j.geoderma.2023.116740.

[17] Y. Wang, M. Khodadadzadeh, and R. Zurita-Milla, "Spatial+: A new cross-validation method to evaluate geospatial machine learning models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 121, p. 103364, 2023, doi: 10.1016/j.jag.2023.103364.

[18] S. de Bruin, D. J. Brus, G. B. M. Heuvelink, T. van Ebbenhorst Tengbergen, and A. M. J.-C. Wadoux, "Dealing with clustered samples for assessing map accuracy by cross-validation," *Ecol. Inform.*, vol. 69, p. 101665, 2022, doi: 10.1016/j.ecoinf.2022.101665.

[19] A. Stock, "Choosing blocks for spatial cross-validation: lessons from a marine remote sensing case study," *Frontiers in Remote Sensing*, vol. 6, p. 1531097, 2025, doi: 10.3389/frsen.2025.1531097.

[20] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023, doi: 10.1002/widm.1484.

[21] M. Baratchi *et al.*, "Automated machine learning: past, present and future," *Artif. Intell. Rev.*, vol. 57, no. 5, p. 122, 2024, doi: 10.1007/s10462-024-10726-1.

[22] R. Gomes Mantovani *et al.*, "Better trees: an empirical study on hyperparameter tuning of classification decision tree induction algorithms," *Data Min. Knowl. Discov.*, vol. 38, no. 3, pp. 1364–1416, 2024, doi: 10.1007/s10618-024-01002-5.

[23] Y. Wei, H. Hu, R. Li, X. Li, and F. Wang, "Digital Mapping of Soil and Water Indicators in Arid Regions Driven by High-Dimensional Environmental Covariates: A Comprehensive Evaluation of Metaheuristic

Feature Selection and Hybrid Deep Learning Frameworks," *Remote Sens. (Basel).*, vol. 18, no. 17, p. 2859, 2026, doi: 10.3390/rs18172859.

[24] S. Dasgupta *et al.*, "Developing regional soil micronutrient management strategies through ensemble learning based digital soil mapping," *Geoderma*, vol. 433, p. 116457, 2023, doi: 10.1016/j.geoderma.2023.116457.

[25] Y. Tian *et al.*, "Digital Mapping of Soil pH Using Tree-Based Models Coupled with Residual Kriging," *Land (Basel).*, vol. 15, no. 3, p. 365, 2026, doi: 10.3390/land15030365.

[26] Z. Zhuo *et al.*, "Enhancing Digital Mapping of Soil Organic Matter in Cropland With Bayesian-Machine Learning Synergy," *Soil Use Manag.*, vol. 42, no. 2, p. e70225, 2026, doi: 10.1111/sum.70225.

[27] J. Zhang, N. Hou, H. Yang, X. Han, and M. Huo, "Prediction of soil organic carbon and nitrogen: a WMTR-CatBoost model based on multi-feature fusion," *Eur. J. Remote Sens.*, vol. 59, no. 1, p. 2711851, 2026, doi: 10.1080/22797254.2026.2711851.

[28] S. Tao, P. Peng, Y. Li, H. Sun, Q. Li, and H. Wang, "Supervised contrastive representation learning with tree-structured parzen estimator Bayesian optimization for imbalanced tabular data," *Expert Syst. Appl.*, vol. 237, p. 121294, 2024, doi: 10.1016/j.eswa.2023.121294.

[29] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[30] M. A. Lones, "Avoiding common machine learning pitfalls," *Patterns*, vol. 5, no. 10, p. 101046, 2024, doi: 10.1016/j.patter.2024.101046.

[31] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artif. Intell. Rev.*, vol. 58, no. 11, p. 339, 2025, doi: 10.1007/s10462-025-11326-3.

[32] S. Kapoor *et al.*, "REFORMS: Consensus-based Recommendations for Machine-learning-based Science," *Sci. Adv.*, vol. 10, no. 18, p. eadk3452, 2024, doi: 10.1126/sciadv.adk3452.

[33] S. Friedrich and T. Friede, "On the role of benchmarking data sets and simulations in method comparison studies," *Biometrical Journal*, vol. 66, no. 1, p. 2200212, 2024, doi: 10.1002/bimj.202200212.

[34] J. L. J. Pereira, K. Smith-Miles, M. A. Muñoz, and A. C. Lorena, "Optimal selection of benchmarking datasets for unbiased machine learning algorithm evaluation," *Data Min. Knowl. Discov.*, vol. 38, no. 2, pp. 461–500, 2024, doi: 10.1007/s10618-023-00957-1.

[35] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *J. Am. Stat. Assoc.*, vol. 119, no. 546, pp. 1434–1445, 2024, doi: 10.1080/01621459.2023.2197686.

[36] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sens. (Basel).*, vol. 16, no. 3, p. 533, 2024, doi: 10.3390/rs16030533.

[37] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.

[38] T. Viering and M. Loog, "The Shape of Learning Curves: A Review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7799–7819, 2023, doi: 10.1109/tpami.2022.3220744.

[39] J. Schmidinger and G. B. M. Heuvelink, "Validation of uncertainty predictions in digital soil mapping," *Geoderma*, vol. 437, p. 116585, 2023, doi: 10.1016/j.geoderma.2023.116585.

[40] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, 2023, doi: 10.1007/s10994-023-06336-7.

[41] J. V. Ribeiro, S. Barbon Junior, F. R. dos Santos, and F. L. Melquiades, "Explainability in vis-NIR and XRF-based modeling for soil fertility analysis: A comprehensive review," *TrAC Trends in Analytical Chemistry*, vol. 199, p. 118776, 2026, doi: 10.1016/j.trac.2026.118776.

[42] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in Deploying Machine Learning: A Survey of Case Studies," *ACM Comput. Surv.*, vol. 55, no. 6, pp. 1–29, 2023, doi: 10.1145/3533378.