Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 2, No. 1, [June 2023], pp. [11–20]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Feature Selection and Spatial Cross-Validation for Optimizing Extremely Randomized Trees in Relative Soil Fertility Classification

Septiana Vratiwi [1*], Hilmi Aziz Bukhori [2], Siti Nur Nazhifah Binti Jais Meah [3], Sadia Samar Ali [4]

[1] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[2] *Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia*
[3] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[4] *King Abdulaziz University (KAU), Jeddah, Arab Saudi*

[1] *septiana@studentmail.unimap.edu.my;* [2] *hilmibukhori@ub.ac.id;* [3] *sitinurnazhifahjm@gmail.com;* [4] *sadiasamarali@gmail.com*
** Corresponding author*

Abstract

This study evaluates feature-mask and hyperparameter selection for Extremely Randomized Trees in relative soil fertility classification. The local dataset contains 135,143 rows, all with the same original Fertility_Level. A 20,000-row sample is assigned three relative classes using training-only Fertility tertiles at 431 and 485. A coordinate-block split yields 15,747 training and 4,253 test observations. The baseline uses 180 trees, while a 12-trial search evaluates tree count, feature fraction, minimum leaf size and five binary inclusion indicators using three-fold group cross-validation. The selected mask retains Latitude, Longitude, NDVI, Soil_Moisture and Elevation, so no dimensionality reduction is achieved. Recorded test macro-F1 increases from 89.10% to 89.54%, and accuracy from 92.71% to 92.92%. Class-1 recall increases from 76.21% to 78.63%, while class-0 recall decreases. The best stored CV macro-F1 is 85.55%. The limited budget includes ten default random startup trials, and the exact winning tree settings are not embedded. The evidence supports a small configuration-associated gain on one split, rather than stable feature selection, optimizer superiority or validated agronomic classification.

Keywords: relative soil fertility; Extra Trees; 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 evaluates a five-variable binary feature mask and Extra Trees hyperparameters under three-fold group cross-validation. A fixed 180-tree baseline and the selected model are compared on the same coordinate-block holdout. The configured search contains 12 trials. The reported winning mask retains every predictor, so the study evaluates a feature-selection procedure without claiming an achieved reduction in dimension.

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 [9] and recursive variable selection for multiple soil properties [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 [25], and Bayesian–machine-learning synergy for soil organic matter combines spatial modeling with a machine-learning predictor [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 [21], and empirical tree-tuning research demonstrates the importance of stating the effective tree configuration [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.

TPE has been used within imbalanced-tabular model development to select a representation-learning parameter [28]; general optimization guidance places such a search within a declared evaluation budget [20]. The current feature-mask and Extra Trees search contains only 12 trials, and its TPESampler call leaves the random-startup count at its default of ten. Thus, at most two completed proposals can follow startup under this configuration. The mask is selected once, and the code does not implement repeated-mask stability selection or bootstrap feature-frequency analysis. 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 [14], whereas partition-randomness research examines the effect of resampling choices in digital soil mapping [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 [39]; classifier-calibration research similarly distinguishes probability reliability from hard-label correctness [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

Stage	Implemented handling
Inner validation	GroupKFold with three folds; validation sizes 5,253, 5,247 and 5,247
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.

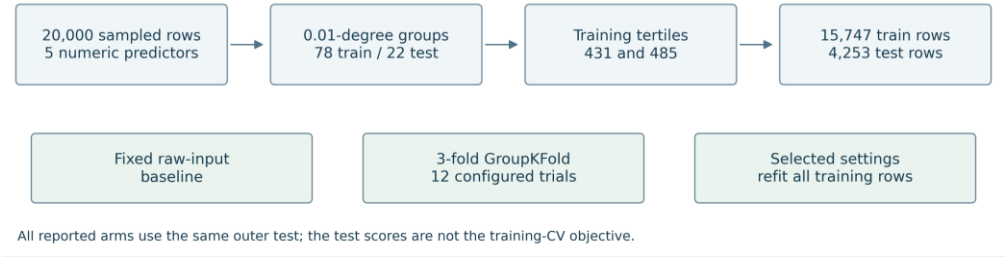

Figure 1. Shared coordinate-block data boundary and comparison arms

ExtraTreesClassifier: n_estimators=180, max_features=sqrt, min_samples_leaf=1, random_state=42 and n_jobs=2; bootstrap remains False and criterion remains gini.

Table 3. Recorded comparison and controlled factors

Component	Study-specific implementation
Fixed comparator	ExtraTreesClassifier: n_estimators=180, max_features=sqrt, min_samples_leaf=1, random_state=42 and n_jobs=2; bootstrap remains False and criterion remains gini.
Selection arm	Binary inclusion indicators for the five original variables, with at least three selected; simultaneous search over tree count, max_features and min_samples_leaf. Twelve trials, TPESampler seed 42 and ten default random startup trials.
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

2.4. Proposed Method

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.

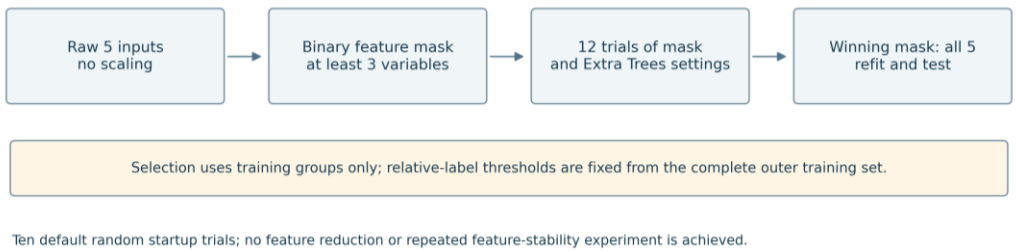

Figure 2. Implemented method and selection boundary

Table 4. Search domain and evaluation controls

Component	Configured domain or control
Feature mask	Five Boolean indicators; configurations retaining fewer than three variables receive objective value zero

Component	Configured domain or control
n_estimators	100–240, step 20
max_features	Continuous fraction 0.50–1.00
min_samples_leaf	Integer 1–4
Fixed settings	random_state=42; n_jobs=2; bootstrap=False; criterion=gini
Objective / budget	Mean three-fold group-CV macro-F1; 12 trials; default ten random startup trials

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 $x \in \mathbb{R}^5$ for the original numeric inputs in the order Latitude, Longitude, NDVI, Soil_Moisture and Elevation. A Boolean mask m selects the subvector in (3). At least three variables must be included; a trial violating this condition returns an objective of zero without fitting the candidate. No scaling or engineered predictors are added.

$$S(m) = \{j : m_j = 1\}, \quad m \in \{0,1\}^5, \quad |S(m)| \geq 3, \quad x^{(m)} = \left(x_j\right)_{j \in S(m)}. \tag{3}$$

For tree b and node t, let n_{bt} be the number of fitting observations reaching the node and n_{btk} the class-k count. With the default Gini criterion, class proportions and impurity are given in (4). Bootstrap is False, so each tree uses the fitting rows without bootstrap resampling. Random feature subsets and randomized thresholds provide tree-level variation.

$$p_{btk} = \frac{n_{btk}}{n_{bt}}, \qquad G_b(t) = 1 - \sum_{k=0}^{2} p_{btk}^2. \tag{4}$$

For a randomized candidate split with children t_L and t_R, (5) gives its local impurity decrease. The admissible candidate with the highest decrease is selected under the minimum-leaf constraints. The max_features parameter controls the considered feature subset; min_samples_leaf controls the minimum observations at a leaf.

$$\Delta G_b(t) = G_b(t) - \frac{n_{bt_L}}{n_{bt}} G_b(t_L) - \frac{n_{bt_R}}{n_{bt}} G_b(t_R). \tag{5}$$

Let $t_b\left(x^{(m)}\right)$ be the leaf reached by an input in tree b. Equation (6) averages the tree probability estimates over B trees and selects the largest class score. This matches mean-probability aggregation rather than a hard majority vote over tree labels. No probability calibration is performed.

$$\hat{p}_k(x) = \frac{1}{B} \sum_{b=1}^{B} p_{b,t_b\left(x^{(m)}\right),k}, \qquad \hat{y}(x) = \arg\max_{k \in \{0,1,2\}} \hat{p}_k(x). \tag{6}$$

Let TP_k, FP_k and FN_k be counts on an evaluation fold. The primary endpoint is the unweighted macro-average in (7); accuracy, weighted F1 and mean class recall are reported additionally.

$$\mathrm{MacroF1}(y, \hat{y}) = \frac{1}{3} \sum_{k=0}^{2} \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}. \tag{7}$$

For the three group-disjoint validation folds $\mathcal{I}_f \subset \mathcal{T}$ and tree settings θ, (8) averages the fold scores. The superscript $-f$ means that the classifier is fitted on the other two folds. The outer-training thresholds in (1) remain fixed. Candidate selection in (9) is restricted to completed trials and admissible masks; it does not optimize the outer-test score.

$$J(m, \theta) = \frac{1}{3} \sum_{f=1}^{3} \mathrm{MacroF1}\left(y_{\mathcal{I}_f}, \hat{y}_{\mathcal{I}_f}^{(-f;m,\theta)}\right). \tag{8}$$

$$(m^*, \theta^*) = \underset{(m,\theta)\in\mathcal{H}_{\text{done}},\ |S(m)|\geq 3}{\arg\max} J(m, \theta). \tag{9}$$

The 12-trial search uses TPESampler with seed 42 and default ten random startup trials. Thus, at most two trials can use adaptive density-based proposals if every configured trial completes. The stored winning mask has all five indicators equal to one. The selected classifier is refitted on all outer-training rows, but its three winning tree hyperparameters are not embedded in the attached output.

The output explicitly lists all five variables as selected and reports best CV macro-F1=0.8555062049544575. It does not print the winning n_estimators, max_features or min_samples_leaf. The referenced best_params.json and result archive are not supplied with the attached notebook, so those three values are not inferred from the holdout scores.

The supplied notebook runs in FAST_MODE and reports one search seed and one outer split. It is used as an archived experiment, without rerunning or selecting additional models from its test scores. Full trial records and a pinned package environment would be needed for a complete reproduction.

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. (%)
ExtraTrees_Baseline	92.71	89.10	92.60	88.64
ExtraTrees_Optimized	92.92	89.54	92.88	89.41

ExtraTrees_Optimized correctly classifies 3,952 test observations, compared with 3,943 for ExtraTrees_Baseline. The differences are 0.21 percentage points in accuracy and 0.45 points in macro-F1. Balanced accuracy increases from 88.64% to 89.41%. The best stored inner-CV macro-F1 is 85.55%, which is a development score on different validation rows rather than another estimate from the outer test.

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
ExtraTrees_Baseline	0	2402	0.9746	0.9900	0.9822
ExtraTrees_Baseline	1	786	0.8308	0.7621	0.7950
ExtraTrees_Baseline	2	1065	0.8846	0.9070	0.8957
ExtraTrees_Optimized	0	2402	0.9817	0.9833	0.9825
ExtraTrees_Optimized	1	786	0.8229	0.7863	0.8042
ExtraTrees_Optimized	2	1065	0.8869	0.9127	0.8996

The selected model correctly predicts 618 class-1 observations, compared with 599 for Baseline. Class-1 assignments to class 0 fall from 62 to 44, and assignments to class 2 fall from 125 to 124. Class-2 correct predictions increase from 966 to 972. These changes are partly offset by class-0 correct predictions falling from 2,378 to 2,362; its misclassification into class 1 rises from 23 to 40. The aggregate gain therefore includes class-specific trade-offs.

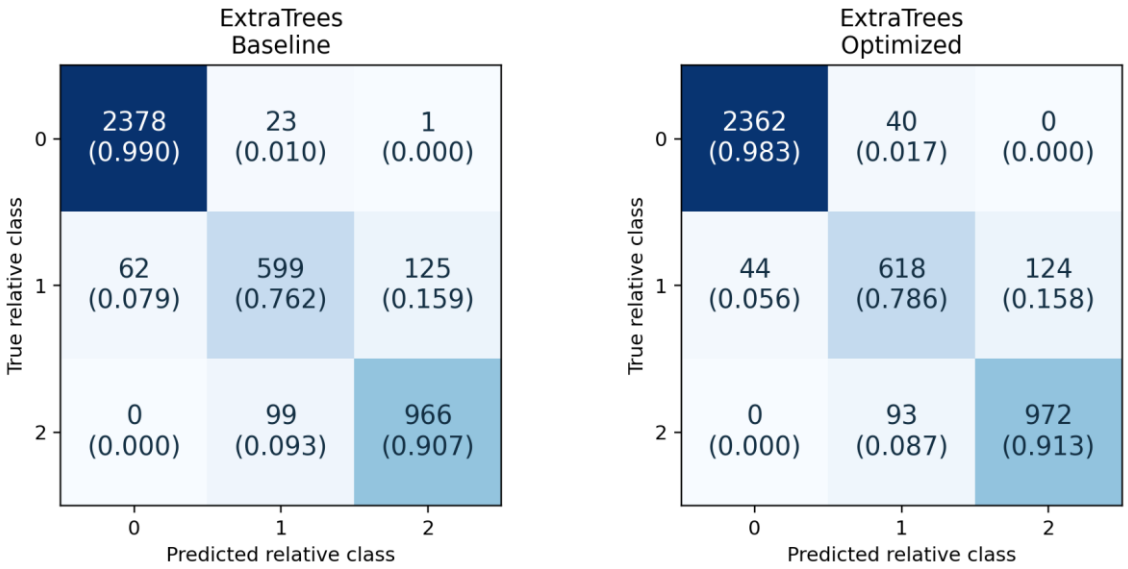

Figure 3. Test confusion matrices; entries show counts and row-normalized proportions

3.3. Internal Learning Curves

Figure 4 reproduces the recorded internal training and validation curves. Fitting subsets contain 3,020, 6,041, 9,062 and 12,083 observations; the internal validation partition contains 3,664 rows. The split is made inside the outer training data using GroupShuffleSplit, with test_size=0.23 and random_state=47. Smaller fitting subsets are stratified row samples from the internal fitting pool. Selected settings remain fixed during curve refits.

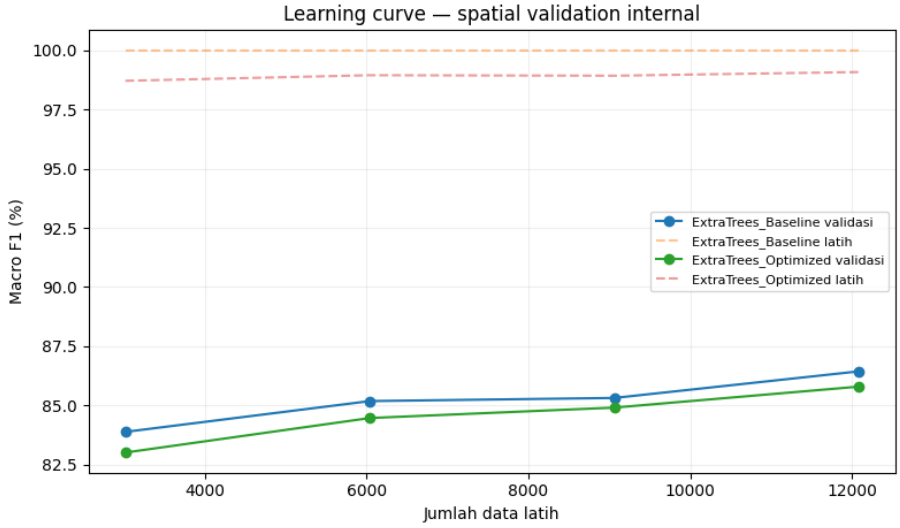

Figure 4. Archived training and internal-validation macro-F1 learning curves

Both internal validation curves rise with fitting size, but the selected model remains below Baseline at every displayed point. Baseline training macro-F1 is essentially 100%, while the selected model is slightly lower and both validation curves remain substantially lower. The small outer-test advantage of the selected model therefore is not reproduced as a uniform advantage on this reused internal partition.

3.4. Discussion and Limitations

The retained mask contains all five original variables. Consequently, the results do not demonstrate a benefit from removing irrelevant predictors, a more compact representation or stability of a selected subset. The observed contrast combines the bounded search and any changed Extra Trees settings. Since only the selected column list and CV score are embedded, a separate feature-mask effect and a precise hyperparameter attribution cannot be recovered. 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. These limits require a bounded interpretation of the observed configuration differences.

4. Conclusion

The recorded Extra Trees selection increases outer-test macro-F1 from 89.10% to 89.54% and accuracy from 92.71% to 92.92%. Class-1 and class-2 recall improve, while class-0 recall decreases. All five inputs are retained, so the experiment achieves no feature reduction. The 12-trial budget includes ten default random startup trials, and the exact winning tree parameters are absent from the attached record. These results describe a small configuration-associated difference on one coordinate-block split. Future work should archive the full winning settings, test feature-mask stability, repeat selection under multiple spatial partitions and validate the relative target against independent soil measurements.

References

[1] M. Sujatha and C. D. Jaidhar, "Machine learning-based approaches to enhance the soil fertility—A review," *Expert Systems with Applications*, vol. 240, Art. no. 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," *Communications in Soil Science and Plant Analysis*, 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, Art. no. 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 – Soil, Air, Water*, vol. 53, no. 9, Art. no. 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, Art. no. 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," *International Journal of Remote Sensing*, 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, Art. no. 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, Art. no. 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, Art. no. 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, Art. no. 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, Art. no. 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," *Applied Soft Computing*, vol. 133, Art. no. 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, Art. no. 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, Art. no. 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, Art. no. 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, Art. no. 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," *Ecological Informatics*, vol. 69, Art. no. 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, Art. no. 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, Art. no. e1484, 2023, doi: 10.1002/widm.1484.
[21] M. Baratchi et al., "Automated machine learning: past, present and future," *Artificial Intelligence Review*, vol. 57, no. 5, Art. no. 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 Mining and Knowledge Discovery*, 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 Sensing*, vol. 18, no. 17, Art. no. 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, Art. no. 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*, vol. 15, no. 3, Art. no. 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 and Management*, vol. 42, no. 2, Art. no. 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," *European Journal of Remote Sensing*, vol. 59, no. 1, Art. no. 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 Systems with Applications*, vol. 237, Art. no. 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, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[30] M. A. Lones, "Avoiding common machine learning pitfalls," *Patterns*, vol. 5, no. 10, Art. no. 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," *Artificial Intelligence Review*, vol. 58, no. 11, Art. no. 339, 2025, doi: 10.1007/s10462-025-11326-3.
[32] S. Kapoor et al., "REFORMS: Consensus-based Recommendations for Machine-learning-based Science," *Science Advances*, vol. 10, no. 18, Art. no. 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, Art. no. 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 Mining and Knowledge Discovery*, 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?," *Journal of the American Statistical Association*, 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 Sensing*, vol. 16, no. 3, Art. no. 533, 2024, doi: 10.3390/rs16030533.
[37] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Transactions of the Association for Computational Linguistics*, 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 Transactions on Pattern Analysis and Machine Intelligence*, 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, Art. no. 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," *Machine Learning*, 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, Art. no. 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 Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2023, doi: 10.1145/3533378.