Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 2, No. 2, [December 2023], pp. [75–86]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Tree-Structured Parzen Estimator–Optimized Decision Trees with Local Spectral Contrast Descriptors for Multiclass Landsat Classification

Mazlina Abdul Majid[1*], Septiana Vratiwi[2], Ronal Watrianthos[3], Hilmi Aziz Bukhori[4]

[1] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pahang, Malaysia.*
[2] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[3] *Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia.*
[4] *Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia*

[1] *mazlina@umpsa.edu.my;* [2] *septiana@studentmail.unimap.edu.my;* [3] *ronalwatrianthos@pnp.ac.id;* [4] *hilmibukhori@ub.ac.id*
** Corresponding author*

Abstract

Multispectral land-cover classification requires models that distinguish spectrally overlapping categories without overstating gains from feature engineering and parameter search. This study compares a raw-feature decision tree, an untuned contrast-feature tree, and a search-selected contrast-feature tree on the UCI Statlog Landsat Satellite benchmark. Sixteen local descriptors expand the 36 original measurements to 52 predictors. The official split contains 4,435 training and 2,000 held-out test records. Ten configurations are evaluated through Optuna's TPESampler using a stratified training-only validation split. With the documented default of ten startup trials, this budget remains within random startup sampling and does not demonstrate model-guided TPE proposals. The baseline achieves 84.90% accuracy and 83.04% macro-F1; the untuned contrast model achieves 84.10% and 82.61%; the selected model achieves 86.40% and 84.77%. The selected configuration improves accuracy by 1.50 percentage points and macro-F1 by 1.73 points relative to baseline. Class-4 recall increases from 55.45% to 64.45%, while some other classes lose recall. The results support a descriptive improvement for the selected configuration on this fixed split, rather than a demonstrated advantage of model-guided TPE. Additional seeds, a search budget extending beyond startup, and geographically separate evaluation are needed before broader conclusions.

Keywords: Decision tree; Hyperparameter optimization; Landsat classification; Local spectral contrast; Tree-structured Parzen estimator.

1. Introduction

Accurate classification of multispectral Landsat land cover requires evaluations that distinguish useful predictive patterns from performance induced by a particular training sample [1], [2]. Recent surveys of satellite mapping identify variation in training observations, representation and geographic conditions as practical limitations [3], [4], [5]. The present study therefore treats Decision Tree as a controlled alternative to a specified baseline rather than assuming that additional model complexity necessarily improves reliability [6], [7]. Environmental and remote sensing features can be informative, but their utility changes with sensor choice, local covariates and the prediction target [8], [9]. In this experiment, every comparison uses the same recorded notebook split and the same class definitions. This is essential because tuning performance and final performance answer different scientific questions: selecting a configuration and estimating its behavior on unseen observations [10], [11]. The resulting comparison is descriptive rather than a confidence interval for an entire geographic population.

This motivates explicit comparison of predictor representations under the same development and evaluation boundary instead of attributing all changes in accuracy to an optimizer [1], [8], [12], [13]. Modern nonlinear learners can extract complex interactions, while compact feature learners and targeted feature engineering provide alternative representations [14], [15], [16], [17]. Hyperparameter optimization can systematically compare candidate

configurations, but its outcome is sensitive to search ranges, evaluation metric, trial budget and resampling strategy [18], [19]. Using the development set to select hyperparameters and then reporting the same validation score as the final evidence risks overly optimistic model assessment [10], [20].

Probability calibration and class-specific behavior are distinct from aggregate classification accuracy and should not be inferred from accuracy alone [21], [22], [23]. Spatially structured observations require special care because samples close in location or other dependencies may provide less independent evidence than their count suggests [24], [25], [26]. Performance on a geographically distinct holdout is a targeted test of transfer within the available sampling frame, not a demonstration of universal geographic validity [24], [27]. The Statlog benchmark contains 3×3 multispectral patches, which allow descriptors of local variation but do not supply the scene coordinates needed for geographical block validation. The recorded official split is therefore treated as a benchmark boundary rather than evidence of independent-scene generalization.

Existing Landsat studies illustrate the value of feature relations, local textures and explicit optimization comparisons without establishing the particular result reported here [1], [8], [28]. The specific research question is whether spectral contrast descriptors and a search-selected decision tree change held-out Macro F1 and class-wise errors when evaluated against the matched baseline. A meaningful comparison requires both an absolute comparison with the default algorithm and separation of the contribution of any additional feature representation [6], [18], [29]. The experimental design reports class-level precision, recall, F1 and support with confusion heatmaps and learning curves, while avoiding claims of statistical significance absent repeated independent trials [22], [30], [31].

The motivation for selecting Decision Tree in this study is not to introduce another algorithm into an already crowded benchmark, but to examine a particular source of model behavior under a traceable experimental protocol. The classifier receives 36-channel 3×3 patch vector, and its decisions are evaluated with respect to six original land-cover categories. These details are not interchangeable: altering the input definition or the target formation changes the learning problem itself. Consequently, the paper identifies the dataset version, feature construction, evaluation unit, and exact performance measure before interpreting a score difference. Such specification also makes the experiment falsifiable, because a future implementation could reproduce each comparison with the archived training and testing files. Transparent specification of the data boundary and comparator also supports meaningful benchmarking [29], [30], [32].

The central methodological intervention concerns local contrast descriptors. This component offers a concrete hypothesis: the model may exploit predictor relationships or parameter configurations that the original comparator does not use as effectively. That hypothesis does not imply that additional representation complexity is always beneficial. More constructed predictors may introduce redundancy [12], [13], while more flexible search may select a configuration that is unusually favorable to its validation subset [11], [20]. The analysis therefore examines the proposed arm against the contrast-only arm, rather than presenting the final tuned model as the sole outcome of the investigation. This distinction matters because a method can appear improved when benchmarked against a weak reference even though a simpler alternative performs just as well [29], [32].

A further concern is the operational meaning of a baseline. For Decision Tree, the baseline must be a specific, executable configuration trained on the same observations and evaluated with the same class labels. It should not be replaced retroactively by a model with different predictors, extra training data, or a different test set. The held-out partition of this study is fixed as the official 2,000-case UCI test set. Comparisons made under that boundary can isolate observed test-set differences; they cannot by themselves quantify performance variability across independent acquisitions or new environmental domains. This is why the manuscript reports the baseline, proposed result, and intermediate ablation where available, with model names matching the saved notebook outputs. Comparator identity and a declared evaluation boundary are central to reproducible comparison [30], [32].

The search procedure uses ten trials through Optuna’s TPESampler, a deliberately limited computational budget that allows the supplied notebook to execute without an extensive infrastructure requirement. An adaptive optimizer can allocate later trials toward promising configurations, but this general capability depends on its search space, objective, initialization and seed [18], [19]. The actual startup limitation of the ten-trial run is stated in Section 2.4. The selected candidate is described as the best configuration among the attempted trials, not the mathematically optimal member of the entire model family][6], [11]. Repeating an experiment with a larger budget may change the selected parameters and the ranking. The objective of this paper is to document the current observable effect on held-out Macro F1 while acknowledging the difference between a successful limited search and a robust conclusion of global superiority.

For multiclass prediction, reporting only accuracy would obscure the error distribution among the six original land-cover categories [22], [33]. A model can improve correct classification of a prevalent category while reducing recall for a smaller or spectrally ambiguous group. The study therefore reports complete precision, recall, F1-score and support for each class, together with macro and weighted averages. The principal distinction is that Macro F1 gives equal contribution to the performance of each class, whereas weighted metrics reflect their support frequencies [22], [23]. This choice is especially important when a new feature set shifts errors rather than uniformly resolving them. A class-level table and separate confusion heatmaps for the reference and proposed model make those trade-offs inspectable.

The two types of confusion-matrix representation serve complementary purposes. Absolute cell counts allow the reader to see how many examples were assigned correctly or incorrectly, while row-normalized values reveal the conditional distribution of predicted labels for a true class [22], [23]. This information supports a more precise discussion of error migration: for example, one model may reduce confusion between two neighboring labels but increase another type of mistake. The numerical summary in the main model-comparison table remains necessary, but the visual evidence prevents an overgeneralized interpretation of a small improvement [33], [34]. Each figure is therefore introduced and discussed separately rather than placed immediately next to another figure without analysis.

Learning-curve analysis provides another perspective on the recorded model behavior. Training and validation scores obtained at increasing development-sample sizes help identify whether the algorithm continues to benefit from additional data under the same feature and optimization configuration [31]. Nevertheless, the curve must be interpreted cautiously because validation subsets are drawn from the development population, not from independently collected geographic regions or future observations. A widening train–validation gap would warrant a different methodological response from two curves that converge at a moderate level. In either case, the curve supplies context rather than an independent confidence interval for the held-out result [31], [35]. The article uses it to discuss modeling constraints without claiming empirical stability that the notebook has not measured.

The benchmark is also limited in the kind of independence it can establish. The Statlog Landsat data encode neighborhoods as 3×3×4 vectors, but do not contain the original geographic scene, coordinates or dates needed to construct verifiable geographic transfer tests. Thus the official test set is a legitimate benchmark boundary without being a substitute for spatially disjoint satellite scenes [24], [25], [26]. The proposed local feature transformation must not be described as if it introduces an additional georeferenced image or a new sensing modality. This paper distinguishes information extracted from an existing patch from information obtained through external spatial measurements, and it recommends scene-separated evaluation whenever suitable data become available [25], [36].

Together, these considerations define the contribution and its limits. The manuscript makes a controlled comparison of Decision Tree using the intervention local contrast descriptors, links the actual implementation to the Method diagrams and parameter table, and reports both global scores and class-specific mistakes. It does not infer statistical significance from one held-out partition, nor does it present increased model complexity as a desirable outcome in its own right [29], [30]. The decisive question is whether the proposed component leads to a useful improvement that remains visible when the intermediate configuration and strongest available comparator are considered. By documenting both favorable and unfavorable outcomes, the study offers a reproducible foundation for a larger multi-seed, cross-domain experiment.

2. Method

2.1. Related Work

Current satellite-mapping research links classification quality to the representation, provenance and validation of training observations [2], [3], [4], [5]. Landsat feature-relationship models and land-use studies using vegetation indices and texture descriptors illustrate how engineered predictors may reveal distinctions that raw channels alone do not expose [1], [8]. These studies motivate an explicit representation comparison, but their results do not establish that the present local descriptors improve Statlog. Recent explainable and optimized multisensor land-cover research likewise provides a precedent for configuration search under richer input settings [37]. The present experiment remains a fixed patch-record benchmark rather than a full-scene mapping system.

Spectral–spatial learning offers several alternatives to fixed local statistics. Relation-guided optical fusion, contextual convolutional transformers and graph-transformer fusion learn dependencies among channels and nearby observations [9], [14], [15]. Compact attention networks and recent feature-extraction reviews describe further

approaches to learned representation [16], [17], while polynomial spectral-curve modeling addresses nonlinear spectral structure through a different representation [38] . Symbolic approaches also retain an important role: decision trees and random forests have been evaluated for mangrove mapping [39], and modal decision trees extend interpretable models with explicit spatial reasoning [40]. Neither a modal tree nor a neural architecture is implemented here; these papers provide context for comparing fixed local descriptors with raw patch inputs.

Hyperparameter optimization literature emphasizes that search objectives, bounds, initialization, computational budget and validation strategy affect the selected model [18], [19]. Direct evidence for decision-tree tuning is provided by an empirical comparison of CART and C4.5 induction algorithms across multiple datasets [6]; classical-classifier reviews also describe the methodological characteristics of tree models [7]. Satellite-image classification studies have examined hyperparameter-dependent validation outcomes [28]. Other work considers multiobjective optimization with fairness and energy objectives [41], while an environmental modeling study applies TPE to ensemble-learner hyperparameters [42]. These precedents support declaring the search design rather than importing their performance claims. The present run uses a single tree and a ten-trial budget, whose startup limitation is specified in the implemented method.

Reliable method comparison requires separation of configuration selection from final assessment. Reviews of leakage and common machine-learning pitfalls describe how preprocessing or repeated selection can compromise evaluation boundaries [10], [11], [20]. Reproducibility recommendations additionally call for explicit model, data and evaluation specifications [30]. Benchmark-selection research and methodological discussions show why conclusions depend on the chosen datasets and comparator design [29], [32]. Cross-validation estimates also require a clearly stated target of inference [35]. These findings motivate retaining the raw-feature baseline, the untuned contrast model and the selected contrast model as separate arms rather than attributing every observed score change to the sampler. Multiclass assessment research distinguishes overall accuracy, class-wise precision and recall, macro averages and support-weighted summaries [22], [23]. Image-analysis validation studies further identify metric-dependent failure modes and recommend selecting measures that match the evaluation task [33], [34]. These principles motivate the complete per-class table and both count-based and row-normalized confusion matrices. Probability calibration is a separate property of predictive probabilities [21]; local conformal uncertainty mapping provides another reliability perspective in land-cover classification [43]. Neither calibration nor conformal prediction is evaluated in this study. Learning-curve reviews support inspecting sample-size effects [31], while such curves remain internal diagnostics rather than independent replications of a held-out result.

Geospatial validation studies show why random partitions may not represent independent geographic evidence when observations are spatially dependent [24], [25], [26]. A remote-sensing roof-detection study illustrates how the evaluation arrangement can lead to overly favorable accuracy assessments [36]. Cross-domain hyperspectral classification offers a different strategy for testing transfer across data domains [27]. These findings support a limited interpretation of the official Statlog split: its records provide local patch measurements, but the available notebook does not reconstruct geographically separate training and test scenes. The comparison therefore evaluates the recorded representations and configurations under a declared benchmark boundary, leaving multi-scene and repeated-seed validation for future experiments.

2.2. Dataset

The UCI Statlog Landsat Satellite benchmark (https://archive.ics.uci.edu/dataset/146/statlog+landsat+satellite) contains 4,435 official training records and 2,000 official test records, each encoding 36 spectral measurements in a 3×3×4 patch. Six labels (1, 2, 3, 4, 5 and 7) are observed. Official sat.trn is used for model training and parameter selection, while sat.tst is reserved for final evaluation. The source provides no reconstructable geographic scene or geographic block identifiers. Table 1 describes the executed data protocol.

Table 1. Dataset, label and partition specification used in the experiment

Dataset attribute	**Study-specific protocol**
Dataset	UCI Statlog Landsat Satellite
Prediction target	Six original observed land-cover labels
Training/test	4,435 official train / 2,000 official test
Input variables	36 raw spectral values; 52 predictors after appending 16 local descriptors
Evaluation boundary	Official UCI training/test split
Source and scope	Statlog encodes 3×3×4 pixel neighborhoods but provides no recoverable scene coordinates or geographic block identifiers.

As Table 1 shows, the original 36-value patch vector is retained in every experimental arm. The contrast arms append 16 deterministic descriptors, producing 52 inputs. The six observed labels remain 1, 2, 3, 4, 5 and 7. Within sat.trn, a stratified 75:25 split with seed 42 yields 3,326 fitting records and 1,109 validation records for configuration selection. The selected model is then refitted on all 4,435 official training records before evaluation on the 2,000 official test records. Table 2 presents the actual experimental controls and evaluation stages, rather than a prospective design.

Table 2. Experimental data flow and leakage-prevention controls

Stage	Required handling
A. Audit	Check original categories, inputs, class counts and notebook outputs
B. Holdout	Official sat.trn: 4,435 records; official sat.tst: 2,000 records
C. Search	3,326 / 1,109 stratified internal split; ten TPESampler startup trials; seed 42
D. Final fit	Refit each final tree on all 4,435 official training records
E. Test	Official-test metrics and confusion matrices; training-only learning curves

Feature construction is performed separately for each patch and uses no class labels or fitted scaling statistics. The baseline uses the 36 raw measurements; both contrast arms use 52 unscaled predictors. Configuration search uses the internal stratified split of sat.trn, and the official sat.tst records are excluded from the objective. Learning curves also use only sat.trn and reuse the same validation partition. These are internal holdout evaluations, rather than geographic-block cross-validation.

2.3. Research Framework

Figure 1 summarizes the three executed arms: DT_Baseline on 36 raw inputs, Contrast_Untuned on 52 inputs, and Contrast_TPE_DT on the same 52 inputs with selected tree parameters. All three final models use the same official training and test records.

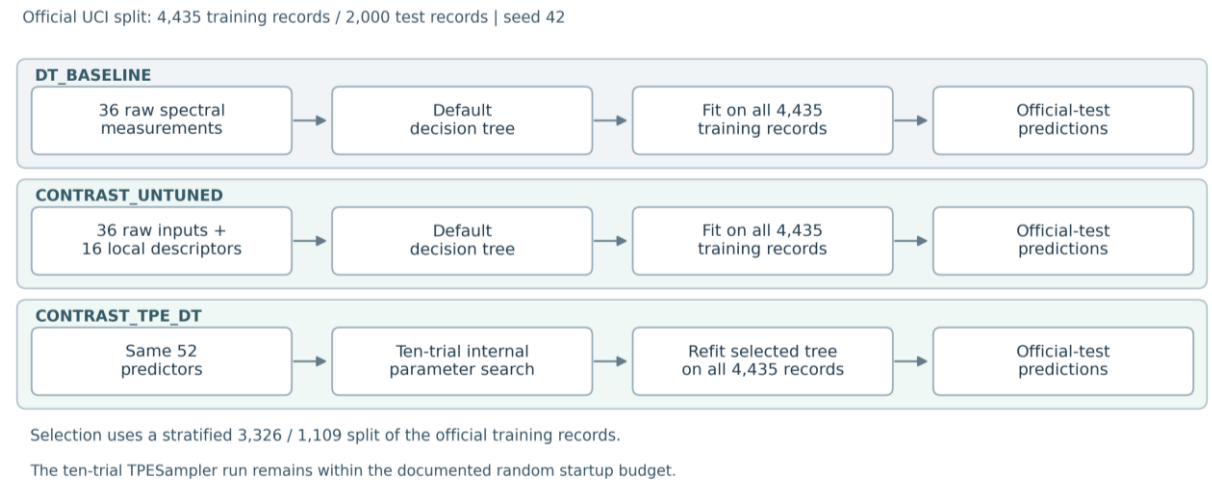


Figure 1. Baseline, local-contrast ablation and parameter-selection framework

As illustrated in Figure 1, the contrast-only arm separates representation from parameter selection. The internal search uses 3,326 training records and 1,109 validation records, followed by refitting on all 4,435 official training records. The common 2,000-record official test set supplies the final model comparison. No repeated independent test replications are reported. Table 3 distinguishes the baseline, optimized configuration and any intermediate ablation actually evaluated in the executable notebook.

Table 3. Controlled baseline, contrast-feature and parameter-selection comparison

Component	Comparative specification
Baseline	DT_Baseline: default decision tree on 36 raw inputs
Proposed	Contrast_TPE_DT: selected decision tree on 52 inputs
Shared evaluation	Same class mapping and held-out observations
Ablation	Contrast_Untuned: default decision tree on the same 52 inputs
Primary assessment	Held-out Macro F1, accuracy, class-wise precision/recall/F1

Table 3 is an executed experimental comparison: all metric differences reported in Section 3 were computed on the same held-out observations. Any extension to additional tuning strategies requires new experiments.

2.4. Proposed Method

The implemented method combines deterministic local spectral descriptors with a search-selected DecisionTreeClassifier. Reshaping each 36-value record to a 3×3×4 array yields four center–neighbor contrasts, four within-patch standard deviations, four horizontal gradients and four vertical gradients. These 16 descriptors are appended to the original 36 values, producing 52 predictors. The standard deviations use all nine pixels; each center contrast subtracts the mean of the eight surrounding pixels from the center value. Horizontal gradients subtract the left-column mean from the right-column mean, and vertical gradients subtract the top-row mean from the bottom-row mean. Figure 2 depicts this exact feature construction and selection sequence.

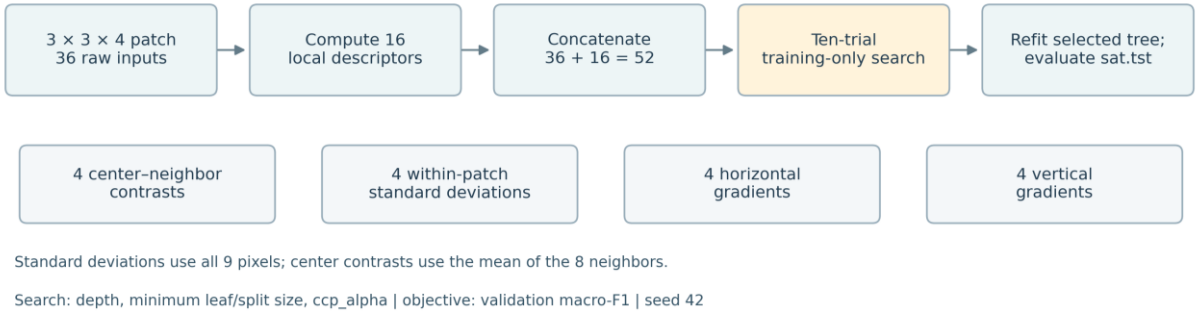


Figure 2. Local spectral descriptors and training-only parameter-search pipeline

Figure 2 shows that the raw-feature and contrast-only trees use DecisionTreeClassifier(random_state=42) with default hyperparameters. The selected contrast model searches max_depth from 6 to 28, min_samples_leaf from 1 to 12, min_samples_split from 2 to 26, and ccp_alpha from 10^{-7} to 0.003 on a logarithmic scale. The impurity criterion is not searched. Optuna maximizes validation macro-F1 over ten trials with TPESampler(seed=42). The notebook does not set n_startup_trials; under the documented default of ten, this run stays within random startup sampling and does not reach model-guided TPE proposals.

For one patch, let X_{rcb} denote the value at row r, column c and band b, with $r, c \in \{1,2,3\}$ and $b \in \{1,2,3,4\}$. Equation (1) defines the within-patch band mean and population standard deviation using all nine positions. The mean is used to compute the standard deviation and is not appended as a separate descriptor.

$$\begin{aligned} \mu_b &= \tfrac{1}{9}\textstyle\sum_{r=1}^{3}\sum_{c=1}^{3} X_{rcb}\,, \\ \sigma_b &= \sqrt{\tfrac{1}{9}\textstyle\sum_{r=1}^{3}\sum_{c=1}^{3}(X_{rcb}-\mu_b)^2}\,. \end{aligned} \tag{1}$$

Let $\mathcal{N} = \{(r,c): r,c \in \{1,2,3\}, (r,c) \neq (2,2)\}$ be the eight surrounding positions. The center contrast is given by (2), and the horizontal and vertical gradients by (3). Their signs follow the right-minus-left and bottom-minus-top definitions used in the implementation.

$$\Delta_b = X_{22b} - \tfrac{1}{8}\textstyle\sum_{(r,c)\in\mathcal{N}} X_{rcb}. \tag{2}$$

$$\begin{aligned} G_b^h &= \tfrac{1}{3}\textstyle\sum_{r=1}^{3}(X_{r3b}-X_{r1b})\,, \\ G_b^v &= \tfrac{1}{3}\textstyle\sum_{c=1}^{3}(X_{3cb}-X_{1cb})\,. \end{aligned} \tag{3}$$

Stacking the four bands of each descriptor gives the representation in (4), where $x \in \mathbb{R}^{36}$ is the original patch vector. Its dimension is $36 + 4(4) = 52$. The tree receives these values directly, without standardization.

$$\phi_c(x) = [x^T, \Delta^T, \sigma^T, (G^h)^T, (G^v)^T]^T \in \mathbb{R}^{52}. \tag{4}$$

For a tree node t, let n_t be its fitting-sample count and n_{kt} the number of those samples in class $k \in \mathcal{L} = \{1,2,3,4,5,7\}$. The class fraction and default Gini impurity are defined in (5). If a split produces children t_L and t_R, its weighted impurity decrease is (6), where N is the total fitting-sample count.

$$p_{kt} = \frac{n_{kt}}{n_t}, \qquad G(t) = 1 - \textstyle\sum_{k\in\mathcal{L}} p_{kt}^2. \tag{5}$$

$$D(t;j,s) = \frac{n_t}{N}\Big[G(t) - \frac{n_{t_L}}{n_t}G(t_L) - \frac{n_{t_R}}{n_t}G(t_R)\Big]. \quad (6)$$

At an eligible node, the greedy split in (7) tests transformed feature j against threshold s. The feasible set $\mathcal{A}_t$ contains the candidate splits allowed by the selected maximum depth d, minimum split size m_{split} and minimum leaf size m_{leaf}: the node depth is below d, $n_t \geq m_{\text{split}}$, and both resulting children contain at least m_{leaf} samples. A node with no admissible split remains a leaf.

$$(j^*, s^*) = \arg\max_{(j,s)\in\mathcal{A}_t} D(t;j,s). \quad (7)$$

The selected complexity parameter $\alpha = \text{ccp_alpha}$ controls pruning through (8). Here $\mathcal{T}_{\text{leaf}}$ is the set of terminal nodes and $|\mathcal{T}_{\text{leaf}}|$ its size. The sample-weighted Gini term is balanced against the number of leaves along the cost-complexity pruning path. For the reached leaf $t(x)$, classification uses its largest class fraction as in (9).

$$R_\alpha(T) = \sum_{t\in\mathcal{T}_{\text{leaf}}} \frac{n_t}{N}G(t) + \alpha|\mathcal{T}_{\text{leaf}}|. \quad (8)$$

$$\hat{y}(x) = \arg\max_{k\in\mathcal{L}} p_{k,t(x)}. \quad (9)$$

Table 4 lists only implementation commitments that can be traced to the code and saved outputs. It does not imply unexecuted nested validation, repeated seeds or calibrated probability testing.

Table 4. Proposed methodological components and verification rules

Design element	Operational definition
Data handling	4,435 official train / 2,000 official test; internal 3,326 / 1,109 split
Core method	36 raw inputs + 4 center contrasts + 4 standard deviations + 4 horizontal + 4 vertical gradients = 52
Tuning	max_depth: 6–28; min_samples_leaf: 1–12; min_samples_split: 2–26; ccp_alpha: 10^{-7}–0.003 (log)
Search budget	Ten trials with TPESampler(seed=42); default ten-trial random startup is not exceeded
Objective	Maximize macro-F1 on the 1,109-record internal validation subset
Ablation	Contrast_Untuned
Evaluation	Official test

The validation criterion is the six-class macro-F1 in (10), where TP_k, FP_k and FN_k are the class-specific validation counts. It gives each observed class equal weight.

$$\text{MacroF1} = \frac{1}{6}\sum_{k\in\mathcal{L}} \frac{2\text{TP}_k}{2\text{TP}_k + \text{FP}_k + \text{FN}_k}. \quad (10)$$

Let $\theta = (d, m_{\text{leaf}}, m_{\text{split}}, \alpha)$ and let $\mathcal{H}_{10}$ contain the ten evaluated configurations. Their domains are $d \in \{6, \ldots, 28\}$, $m_{\text{leaf}} \in \{1, \ldots, 12\}$, $m_{\text{split}} \in \{2, \ldots, 26\}$ and $\alpha \in [10^{-7}, 0.003]$, with logarithmic sampling for α. Equation (11) selects the best evaluated configuration using internal validation predictions $\hat{y}_{\text{val}}^{(\theta)}$ from the corresponding fitting-subset tree. This finite candidate selection does not imply that the ten-trial run reached adaptive TPE sampling.

$$\theta^* = \arg\max_{\theta\in\mathcal{H}_{10}} \text{MacroF1}\Big(y_{\text{val}}, \hat{y}_{\text{val}}^{(\theta)}\Big). \quad (11)$$

Table 4 defines macro-F1 as the selection objective. The recorded best configuration is max_depth=9, min_samples_leaf=2, min_samples_split=3, and ccp_alpha=0.00075049862095636, with internal validation macro-F1 of 0.830628344610198. These parameters are frozen before refitting on all official training records. The validation score is configuration-selection information, distinct from the final official-test macro-F1 of 0.8477.

Implementation and reproducibility. The experiment uses seed 42, FAST_MODE=True, ten parameter trials, a stratified 75:25 training-only split, and fixed official UCI partitions. The stored notebook contains code and outputs but does not record package versions, hardware details, dataset hashes or a full per-trial study log. These items are therefore not claimed as completed measurements. The selected configuration is evaluated descriptively; repeated-seed intervals, independent-scene testing, probability calibration, inference-runtime benchmarks and direct tree-complexity comparisons were not performed.

3. Results and Discussion

3.1. Overall Classification Performance

Table 5 compares all recorded model arms using a common official UCI test set. The principal endpoint is Macro F1, supplemented by accuracy, weighted F1, and balanced accuracy. The proposed model increased by 1.73 percentage points relative to DT_Baseline; the magnitude and direction are interpreted jointly with the intermediate model where available.

Table 5. Comparison of baseline and optimized classifiers on the held-out evaluation set

Model	Accuracy (%)	Macro F1 (%)	Weighted F1 (%)	Balanced Acc. (%)
DT_Baseline	84.90	83.04	84.97	82.93
Contrast_Untuned	84.10	82.61	84.33	82.69
Contrast_TPE_DT	86.40	84.77	86.43	84.78

Table 5 shows that Contrast_TPE_DT achieved 84.77% Macro F1 compared with 83.04% for DT_Baseline, with accuracy increasing from 84.90% to 86.40%. The gains are 1.73 and 1.50 percentage points, respectively. In contrast, Contrast_Untuned attained 82.61% Macro F1 and 84.10% accuracy, below the raw baseline. Parameter selection increased Macro F1 by 2.16 points relative to the untuned contrast model. Thus, adding descriptors alone did not improve the baseline; the selected tree configuration accounts for the recovery in the contrast representation. This ablation does not isolate the advantage of TPE over random search, because the ten-trial run remains within startup sampling.

3.2. Per-Class Performance and Confusion-Matrix Analysis

Table 6 presents full precision, recall, F1 and support for each observed label under DT_Baseline and Contrast_TPE_DT. It enables identification of the lowest-recall classes, which could be obscured by an aggregate metric. Figures 3 and 4 then show absolute and row-normalized confusion heatmaps for the same two final models.

Table 6. Per-class precision, recall, F1 and support for baseline and optimized models

Class	Base P	Base R	Base F1	Opt P	Opt R	Opt F1	Support
1	0.9569	0.9631	0.9600	0.9364	0.9588	0.9475	461
2	0.9731	0.9688	0.9709	0.9817	0.9598	0.9707	224
3	0.8648	0.8539	0.8593	0.8845	0.9068	0.8955	397
4	0.5294	0.5545	0.5417	0.6099	0.6445	0.6267	211
5	0.8414	0.8059	0.8233	0.8103	0.7932	0.8017	237
7	0.8245	0.8298	0.8271	0.8658	0.8234	0.8441	470
macro avg	0.8317	0.8293	0.8304	0.8481	0.8478	0.8477	2000
weighted avg	0.8505	0.8490	0.8497	0.8652	0.8640	0.8643	2000

Table 6 identifies class 4 as the lowest-recall category in the selected model, with recall 0.6445 compared with 0.5545 for the baseline. These class-specific values are more informative than overall accuracy alone, particularly where support differs across classes. Gains are not uniform: recall decreases for classes 1, 2, 5 and 7 while increasing for classes 3 and 4. Figure 3 illustrates the baseline predictions, showing where incorrect assignments are concentrated among the classes.

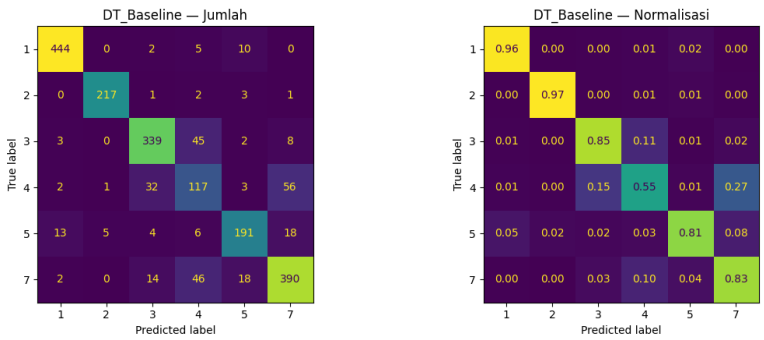


Figure 3. DT_Baseline confusion matrix: absolute counts and row-normalized values

As shown in Figure 3, the baseline made class-dependent errors. Class 4 had 117 correct predictions among 211 observations, corresponding to recall 0.5545. Its largest confusion was with class 7, which received 56 of the class-4 observations. These counts identify the baseline error pattern before comparison with the selected model. Figure 4 displays the optimized model predictions on exactly the same held-out cases.

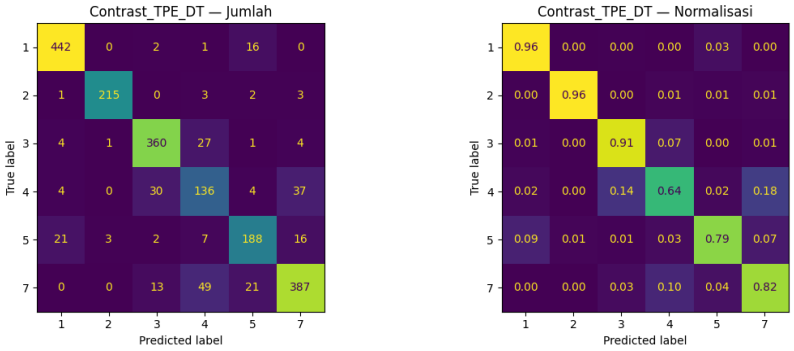


Figure 4. Contrast_TPE_DT confusion matrix: absolute counts and row-normalized values

Figure 4 confirms that 136 of the 211 class-4 observations were correctly predicted, corresponding to recall 0.6445. Misclassification of class 4 as class 7 fell from 56 to 37 observations. However, class-7 recall declined from 0.8298 to 0.8234. The comparison supports a descriptive assessment of changed error patterns rather than statistically significant superiority.

3.3. Learning-Curve Assessment

Figure 5 reports the notebook learning curves for training subsets of 831, 1,663, 2,494 and 3,326 official training records. The 1,109-record internal validation subset remains fixed, and the selected tree parameters are held constant while the models are refitted at each size. The 2,000 official test records are not used to construct these curves. Figure 5 summarizes how the development-sample training and validation scores change with increasing learning-set size.

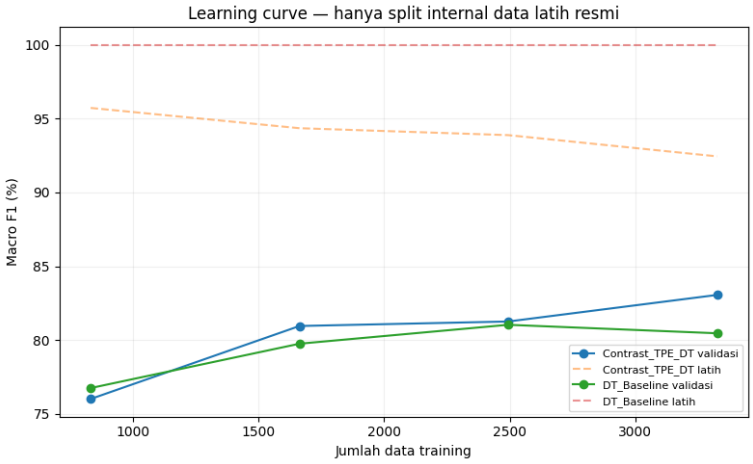


Figure 5. Learning curve recorded in the supplied notebook

Figure 5 shows a persistent training–validation gap. The unrestricted baseline fits its training subsets perfectly, whereas the selected contrast tree has lower training scores and a higher validation score at the largest subset. This is an internal diagnostic, rather than an independent estimate of reproducibility: the selected parameters were chosen using the same validation partition. Additional seeds and geographically separate scenes are required to test whether the pattern persists.

3.4. Discussion and Limitations

The observed gains remain specific to the official holdout and the selected configuration. Feature augmentation alone reduced the baseline scores, so the result should not be attributed to local descriptors in isolation. The ten-trial TPESampler run does not extend beyond the documented startup budget; no comparison against a separate random-

search optimizer or a longer model-guided TPE run was performed. The notebook also does not establish uncertainty intervals, repeated-seed stability, calibrated probabilities or independent-scene transfer. The benchmark provides no usable scene identifiers for geographic blocking of these archived records. These limitations restrict the findings to a descriptive configuration comparison and do not justify statistical significance or an optimizer-specific advantage.

4. Conclusion

The experiment evaluated local spectral contrast descriptors and a search-selected decision tree using the official UCI Statlog training and test partitions. Sixteen descriptors expanded the 36 original inputs to 52. The raw baseline achieved 84.90% accuracy and 83.04% Macro F1; the untuned contrast model achieved 84.10% and 82.61%; Contrast_TPE_DT achieved 86.40% and 84.77%. The selected configuration therefore improved accuracy by 1.50 percentage points and Macro F1 by 1.73 points relative to baseline, while the feature-only ablation was weaker. These outcomes describe the archived configuration and split. Because the ten-trial sampler run remains within documented random startup sampling, the experiment does not establish an advantage of model-guided TPE. Future work should extend the budget beyond startup, record full search and environment details, evaluate repeated seeds and calibration, and test independently verified geographic scenes.

References

[1] P. Dou, H. Shen, C. Huang, Z. Li, Y. Mao, and X. Li, "Large-scale land use/land cover extraction from Landsat imagery using feature relationships matrix based deep-shallow learning," *International Journal of Applied Earth Observation and Geoinformation*, vol. 129, p. 103866, 2024, doi: 10.1016/j.jag.2024.103866.

[2] Y. Wang, Y. Sun, X. Cao, Y. Wang, W. Zhang, and X. Cheng, "A review of regional and Global scale Land Use/Land Cover (LULC) mapping products generated from satellite remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 206, pp. 311–334, 2023, doi: 10.1016/j.isprsjprs.2023.11.014.

[3] S. Zhao, K. Tu, S. Ye, H. Tang, Y. Hu, and C. Xie, "Land Use and Land Cover Classification Meets Deep Learning: A Review," *Sensors*, vol. 23, no. 21, p. 8966, 2023, doi: 10.3390/s23218966.

[4] I. Kotaridis and M. Lazaridou, "Cnns in land cover mapping with remote sensing imagery: a review and meta-analysis," *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 5896–5935, 2023, doi: 10.1080/01431161.2023.2255354.

[5] D. Moraes, M. L. Campagnolo, and M. Caetano, "Training data in satellite image classification for land cover mapping: a review," *Eur. J. Remote Sens.*, vol. 57, no. 1, p. 2341414, 2024, doi: 10.1080/22797254.2024.2341414.

[6] 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.

[7] A. H. Alshammari, G. Bencsik, and A. H. Ali, "A Survey of Six Classical Classifiers, Including Algorithms, Methodological Characteristics, Foundational Variants, and Recent Advances," *Algorithms*, vol. 19, no. 1, p. 37, 2026, doi: 10.3390/a19010037.

[8] Y. Maulani and K. Surendro, "Detailed land use classification model based on vegetation indices and texture features," *Remote Sens. Appl.*, vol. 40, p. 101786, 2025, doi: 10.1016/j.rsase.2025.101786.

[9] X. Geng *et al.*, "A Spatial-Spectral Relation-Guided Fusion Network for Multisource Optical RS Image Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8991–9004, 2025, doi: 10.1109/tnnls.2024.3413799.

[10] 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.

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

[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] 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.

[14] W. Wang *et al.*, "A spatial–spectral fusion convolutional transformer network with contextual multi-head self-attention for hyperspectral image classification," *Neural Networks*, vol. 187, p. 107350, 2025, doi: 10.1016/j.neunet.2025.107350.

[15] Z. Zheng, M. Debbagh, X. Zhou, S. Sun, and Y. Huang, "Graph-Transformer with spatial-spectral features fusion for hyperspectral image classification," *Expert Syst. Appl.*, vol. 264, p. 125962, 2025, doi: 10.1016/j.eswa.2024.125962.

[16] C. Wang *et al.*, "SSFAN: A Compact and Efficient Spectral-Spatial Feature Extraction and Attention-Based Neural Network for Hyperspectral Image Classification," *Remote Sens. (Basel).*, vol. 16, no. 22, p. 4202, 2024, doi: 10.3390/rs16224202.

[17] V. Sangeetha and L. Agilandeeswari, "Artificial intelligence enabled spectral-spatial feature extraction techniques for land use and land cover classification using hyperspectral images – An inclusive review," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 28, no. 3, pp. 455–467, 2025, doi: 10.1016/j.ejrs.2025.06.004.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] X. Zhang and B. Liu, "Open-Set Cross-Domain Hyperspectral Image Classification Based on Manifold Mapping Alignment," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 6241–6252, 2024, doi: 10.1109/jstars.2024.3396883.

[28] F. Alonso-Sarría, C. Valdivieso-Ros, and F. Gomariz-Castillo, "Analysis of the hyperparameter optimisation of four machine learning satellite imagery classification methods," *Comput. Geosci.*, vol. 28, no. 3, pp. 551–571, 2024, doi: 10.1007/s10596-024-10285-y.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] A. Reinke *et al.*, "Understanding metric-related pitfalls in image analysis validation," *Nat. Methods*, vol. 21, no. 2, pp. 182–194, 2024, doi: 10.1038/s41592-023-02150-0.

[34] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nat. Methods*, vol. 21, no. 2, pp. 195–212, 2024, doi: 10.1038/s41592-023-02151-z.

[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] D. Abriha, P. K. Srivastava, and S. Szabó, "Smaller is better? Unduly nice accuracy assessments in roof detection using remote sensing data with machine learning and k-fold cross-validation," *Heliyon*, vol. 9, no. 3, p. e14045, 2023, doi: 10.1016/j.heliyon.2023.e14045.

[37] H. İ. Gündüz, O. Orhan, S. N. G. Hamal, and S. Ekercin, "An Explainable and Optimized Machine Learning Framework for Multisensor Remote Sensing-Based Land Use and Land Cover Classification," *Transactions in GIS*, vol. 30, no. 3, p. e70293, 2026, doi: 10.1111/tgis.70293.

[38] Y. Ruan, Q. Zhao, and J. Sun, "Hyperspectral image classification based on polynomial functional expression of cumulative spectral curve," *Opt. Lasers Eng.*, vol. 194, p. 109167, 2025, doi: 10.1016/j.optlaseng.2025.109167.

[39] A. D. Purwanto, K. Wikantika, A. Deliar, and S. Darmawan, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sens. (Basel).*, vol. 15, no. 1, p. 16, 2023, doi: 10.3390/rs15010016.

[40] G. Pagliarini and G. Sciavicco, "Interpretable land cover classification with modal decision trees," *Eur. J. Remote Sens.*, vol. 56, no. 1, p. 2262738, 2023, doi: 10.1080/22797254.2023.2262738.

[41] A. Candelieri, A. Ponti, and F. Archetti, "Fair and green hyperparameter optimization via multi-objective and multiple information source Bayesian optimization," *Mach. Learn.*, vol. 113, no. 5, pp. 2701–2731, 2024, doi: 10.1007/s10994-024-06515-0.

[42] B.-N. Jiang, Y.-Y. Zhang, Z.-Y. Zhang, Y.-L. Yang, and H.-L. Song, "Tree-structured parzen estimator optimized-automated machine learning assisted by meta–analysis for predicting biochar–driven N2O mitigation effect in constructed wetlands," *J. Environ. Manage.*, vol. 354, p. 120335, 2024, doi: 10.1016/j.jenvman.2024.120335.

[43] D. Valle, R. Leite, R. Izbicki, C. Silva, and L. Haneda, "Local uncertainty maps for land-use/land-cover classification without remote sensing and modeling work using a class-conditional conformal approach," *International Journal of Applied Earth Observation and Geoinformation*, vol. 135, p. 104288, 2024, doi: 10.1016/j.jag.2024.104288.