Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 2, No. 2, [December 2023], pp. [63–74]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Enhancing Logistic Regression for Multispectral Land-Cover Classification Through Polynomial Spectral Interaction Features and Regularization Optimization

Dodi Siregar[1], Siti Nur Nazhifah Binti Jais Meah[2], Muhammad Syahrizal[3], Gerhard Wilhelm Weber[4*]

[1] *Computer Science Program, Faculty of Engineering and Computer Science, Universitas Harapan, Medan, Indonesia*
[2] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[3] *Computer Graphics Engineering Technology Program, Politeknik Cendana, Indonesia*
[4] *Faculty of Engineering Management, Poznan University of Technology, Greater Poland, Poland*

[1] *dodi.stth@gmail.com;* [2] *sitinurnazhifahjm@gmail.com;* [3] *m.syahrizal156@gmail.com;* [4] *gerhard.weber@put.poznan.pl*
** Corresponding author*

Abstract

Multispectral land-cover classification remains difficult when different surface categories exhibit overlapping spectral responses and a linear classifier cannot express interactions among bands. This study investigated whether polynomial spectral interaction features and regularization optimization improve multinomial logistic regression on the UCI Statlog Landsat Satellite dataset. The benchmark provides 4,435 official training cases and 2,000 independent test cases, with 36 spectral attributes per 3 × 3 neighborhood and six observed land-cover classes. Four squared terms and six pairwise products of the four center-pixel bands were added to the original inputs, yielding 46 predictors. Three configurations were compared under the same official test split: a 36-feature logistic-regression baseline, an untuned 46-feature polynomial model, and a 46-feature model with the inverse regularization strength selected by a Tree-structured Parzen Estimator using eight trials on a stratified training-only validation split. The baseline achieved 83.95% accuracy and 79.71% macro-F1; untuned polynomial features increased these scores to 84.20% and 80.41%; the optimized model achieved 85.30% and 82.32%, respectively. Thus, the optimized model improved accuracy by 1.35 percentage points and macro-F1 by 2.61 points relative to baseline. Class-4 recall increased from 30.33% to 40.76%, although its errors remained substantial and class-7 recall declined slightly. These findings show a measurable improvement for the fixed benchmark split while remaining limited by one random seed, eight optimization trials, and the absence of geographically independent evaluation. The method offers a transparent, low-complexity feature-and-regularization ablation that warrants more extensive validation before claims of broader transferability.

Keywords: Landsat land-cover classification; multinomial logistic regression; polynomial spectral interactions; L2 regularization; Tree-structured Parzen Estimator

1. Introduction

Reliable land-cover classification is essential for interpreting agricultural landscapes, monitoring vegetation and urban expansion, and supporting environmental planning from repeated satellite observations [1], [2]. Multispectral instruments such as Landsat offer extensive historical coverage and a manageable number of spectral channels, yet their records remain challenging because different surface materials can display overlapping spectral responses [1], [2], [3]. A single global accuracy score can therefore conceal substantial disagreement for individual cover types, especially mixed vegetation, heterogeneous soil, and transitional terrain [4]. Recent studies of Landsat mapping and large-scale land-cover products show that methodological choices surrounding feature representation and evaluation influence not only benchmark accuracy but also the plausibility of applying a map beyond its development area [1], [2], [3]. Comprehensive evaluations of modern land-cover systems further emphasize differences between experimental classification results and operational reliability, which motivates reporting class-specific errors and not only a single summary number [1], [2], [5]. In this context, classical classifiers deserve rigorous consideration

alongside high-capacity neural architectures: a recent methodological survey compares their interpretability, computational efficiency, decision-boundary complexity, and multiclass handling, including the limitations of linear logistic models under nonlinear effects [6].

Feature representation is a central limitation of linear classifiers when the discriminative information arises from interactions rather than individual bands [6], [7]. Two cover classes may be difficult to distinguish from any one spectral channel but become more separable when the joint response of several bands is considered [3], [8]. Remote-sensing research has addressed this issue using learned feature relations, vegetation indices, local texture descriptors, and spectral–spatial fusion, although the resulting improvements depend on data availability and model capacity [8], [9], [10], [11]. Neural attention and graph-based approaches can learn increasingly complex multiscale relationships, while their greater flexibility increases the need for carefully controlled comparisons [12], [13], [14], [15]. For a compact benchmark, a lower-order polynomial expansion offers a complementary, testable option: pairwise products and squares of selected bands expose nonadditive relationships while retaining a transparent map from the original measurements to the classifier's inputs [7]. Polynomial representations have also been investigated for spectral-curve modeling, supporting the wider relevance of spectral nonlinearity without establishing that the same implementation will improve this Landsat task [16] . Importantly, a feature transformation should not be mistaken for new geographic information; interactions derived from the center pixel in a stored 3 × 3 patch are not substitutes for georeferenced scene context [17].

Multinomial logistic regression is an appropriate methodological anchor for studying these transformations because its softmax decision rule and cross-entropy objective are well defined, its coefficients can be regularized, and its performance can be compared across feature sets under identical training and testing boundaries [6], [16]. Nevertheless, adding interaction terms increases feature dimensionality and may create strong dependencies among predictors [7]. Standardizing features and controlling coefficient magnitude are therefore important for stable numerical estimation and reducing the risk of overfitting [18], [19]. General regularization literature describes coefficient penalization as a way to control model complexity, although its clinical examples do not establish performance gains on this Landsat benchmark [20]. Recent work on automated model selection emphasizes that the effectiveness of regularization depends on the chosen objective, search bounds, validation design, and computational budget; a model selected with extra freedom should not be assumed superior without a held-out assessment [21]. The role of regularization becomes particularly relevant when second-order expansions add predictors derived from the same measurements, because tuning may offset the increased flexibility created by the expansion [7], [18], [20]. Conversely, an improved test result does not establish that the representation and optimization mechanisms interact synergistically unless the intermediate configuration without tuning is also evaluated [22]. Explicit ablation is thus necessary: a raw-feature logistic model, a polynomial-feature model with default regularization, and the same expansion with a validation-selected regularization parameter are compared in the present study.

A separate concern is that regularized logistic regression may present an apparently straightforward model while still producing class-dependent errors that are difficult to interpret from the final accuracy alone [4]. Multinomial outcomes are evaluated through mutually competing class probabilities; changing the coefficient penalty can shift several decision boundaries at once and may improve recall for one class while reducing sensitivity for another [6], [18]. For heterogeneous multispectral patches this trade-off is material, because neighboring spectral signatures need not correspond to labels separated by a wide margin. Research on multiclass assessment shows that class proportions and averaging conventions influence how classification success is summarized [4]. An effective experimental comparison therefore needs complete precision, recall, F1-score, and support, along with both absolute and normalized confusion matrices [4]. The normalized matrix provides a class-conditional perspective when the six categories occur with different frequencies, while the absolute counts retain the magnitude of remaining errors [4]. In addition, learning curves describe how generalization performance changes with training-set size [23]. A curve obtained within the present development sample can reveal a training–validation gap, but it is not an additional independent test because the points share the same development population [23].

Fair evaluation requires more than an optimized score on the data used to choose a parameter [22], [24]. Repeated probing of validation outcomes can make a favorable configuration appear more general than it truly is, and observations collected from nearby sites may not represent independent evidence when there is spatial autocorrelation [17], [21], [24]. Current geospatial validation studies and reviews recommend keeping geographic dependence explicit where group identifiers exist, and distinguish independent test evaluation from selection carried out within the development sample [17], [24]. The Statlog Landsat Satellite dataset provides a fixed official training/test split and 36 numerical predictors formed from four spectral channels at each position in a 3 × 3 patch. It

does not supply the coordinates, full scene mosaic, or location-level group identifiers required for spatial-block cross-validation. Consequently, a stratified random validation split inside the official training data is an available selection strategy, but it cannot establish geographic transferability [17]. An honest experimental description must therefore state which claims the available observations can support, and leave multi-scene transfer, probability calibration, and independently repeated tuning as questions for subsequent studies [17], [22], [25]. Recent research on prediction uncertainty and probability calibration reinforces the distinction between identifying the most likely class and demonstrating trustworthy confidence estimates [25], [26].

The structure of a 3 × 3 Landsat patch also places useful limits on what feature engineering can mean. The four central spectral values are associated with the labeled center pixel, but the 32 remaining values describe surrounding pixels in the archived representation. Keeping these 36 attributes in the raw-feature baseline makes the comparison stronger than replacing them with only four channels: the proposed additional terms must contribute information beyond the original patch vector rather than simply benefit from access to more baseline measurements. Creating interactions of the center bands is computationally inexpensive, yet quadratic terms may be correlated with the original bands and with one another [7]. This is precisely why penalty selection should be evaluated as an explicit additional experimental component rather than assumed to work by default [20], [21]. At the same time, selecting only the inverse regularization strength C with eight TPE trials provides a more limited intervention than jointly optimizing feature subsets, polynomial degree, penalty family, solver, or calibration. That restricted scope makes the comparison reproducible but prevents conclusions about the globally optimal polynomial model or about all possible forms of regularized classification [21], [22].

Spectral–spatial advances provide further motivation to study whether inexpensive explicit interactions can extract part of the benefit obtained with much more elaborate feature learners. Multimodal fusion and transformer-based networks have combined complementary signals to handle complex visual patterns, and cross-domain models aim to improve robustness under distributional shifts [27], [28], [29], [30]. Adaptive spectral–spatial attention, multibranch fusion, and graph reasoning have likewise advanced hyperspectral classification, but such systems typically differ from the present benchmark in their input dimensionality, architecture, and training demands [31], [32], [33], [34]. Few-shot semantic priors and geometric equivariance offer further approaches to hyperspectral representation [35], [36]. Other studies investigate high-order deformable convolutions and semisupervised graph learning [37], [38], segmented-PCA multibranch fusion [39], or multimodal urban information [30]. These research directions show why nonlinear representation matters, but they do not replace a controlled experiment isolating a simple feature expansion under a fixed official split. A comparison limited to classical logistic regression is intentionally narrower than these advanced architectures, and its contribution must be judged within that narrower scope rather than through unsupported claims of state-of-the-art remote-sensing performance.

This study examines whether second-order interactions among the four central spectral bands of each Statlog patch improve multinomial logistic-regression classification, and whether optimizing the regularization strength provides an additional gain. The analysis uses the original 36 measurements as the control representation, adds ten center-band squares and pairwise products to create 46-feature polynomial models, and compares default-regularized and optimized versions on the same 2,000 official test observations. Regularization is selected using a Tree-structured Parzen Estimator (TPE) within a stratified training-only validation split; the polynomial degree is fixed at two, not independently searched. The principal endpoints are macro-F1, overall accuracy, balanced accuracy, and per-class precision and recall, accompanied by confusion-matrix heatmaps and training/validation learning curves. This design responds to the demand for reproducible feature–classifier comparisons with stated search budgets [21], [24] . The intended novelty is a transparent interaction-feature and regularization ablation on this benchmark—not a claim that polynomial expansion, logistic regression, or TPE are individually new. Its scientific value rests on whether a measurable improvement remains after comparison with both the unmodified baseline and the untuned expanded model, while explicitly acknowledging the limits of a single official split and a modest eight-trial optimization budget [22].

2. Method

2.1. Related Work

Current land-cover mapping research connects algorithm performance to the representation and quality of spectral observations [3], [40]. Studies of Landsat feature relationships and physically motivated indices illustrate how engineered predictors can expose class distinctions that may be obscured in raw channels [3], [8]. Reviews of regional mapping products and satellite training data also emphasize that data provenance, annotation design, and

validation choices influence reported model quality [1], [40]. These findings motivate a controlled transformation of available measurements. In the present study, the 36-feature benchmark is a patch-record classification task rather than a full-scene mapping experiment.

Spectral–spatial learning research demonstrates a wide range of alternatives to fixed predictors. Optical-image relation modeling, graph-transformer fusion, multiscale attention, contextual convolutional transformers, and adaptive weighted fusion encode interactions among channels and nearby observations [9], [10], [11], [12], [15]. Multibranch architectures and redundancy-aware networks attempt to combine complementary cues without unnecessarily repeating information [13], [14]. Other investigations examine frequency-domain decomposition [31], multilayer and residual-transformer fusion [32], [33], global attention and graph reasoning [12], [34], and compact attention-based feature extraction [41]. These approaches establish the significance of feature interactions but are not direct experimental baselines for the current work because they assume different architectures, optimization budgets, and sometimes richer input modalities.

Work on generalization and distribution shift introduces additional constraints on how classification gains should be interpreted. Cross-domain alignment, few-shot semantic priors, geometric equivariance, semisupervised graph prediction, and cross-scene optimization aim to improve transfer across sensors, labeled-data regimes, and acquisition conditions [27], [28], [35], [36], [38]. Local uncertainty mapping offers a complementary route for describing reliability beyond mean accuracy [26]. Geospatial validation research emphasizes testing beyond random partitions where location information is available [17]. Such results cannot establish spatial independence for Statlog, whose official record lacks scene coordinates; geographically separated evaluation remains a future extension.

For the present comparison, feature design and optimization are evaluated as separate experimental steps. Hyperparameter search can improve predictive models, but success depends on a declared objective, a feasible parameter range, and an evaluation set not reused for selection [21], [24]. Automated learning frameworks and recent optimized land-cover studies provide relevant precedents for searching model configurations under finite computational budgets [21], [42]. A satellite-classification study comparing RF, SVM, MLP, and CNN examined how hyperparameter choices affect validation accuracy and variability [42]. Multiobjective Bayesian optimization has also considered accuracy, fairness, and energy use [43]; those additional objectives are not evaluated here. Probability calibration research additionally highlights that discrimination metrics such as macro-F1 are not substitutes for calibration or log-loss measurements [25]. The current study therefore limits its conclusions to classification metrics actually produced by the notebook rather than claiming unreported confidence quality, repeated-seed stability, or calibration.

Multisensor fusion, learned hyperspectral descriptors, and recent polynomial spectral-curve representations suggest several ways to model nonlinear relationships, but they do not establish the incremental value of ten simple interaction variables for multinomial logistic regression on Statlog [16], [29], [44]. Different methods of transforming multispectral inputs and tuning machine-learning classifiers have been reported in other remote-sensing contexts [3], [42]. The present comparison resolves a narrower empirical question by evaluating raw 36-feature logistic regression, an untuned 46-feature polynomial version, and a TPE-regularized 46-feature version using exactly the same official test partition. This ablation distinguishes the observed changes following representation and regularization steps while leaving more extensive generalization claims for future work. Hybrid attention–SVM classification offers a different representation-oriented alternative [45], but does not establish the benefit of polynomial logistic regression.

2.2. Dataset

The experiment uses the UCI Statlog Landsat Satellite dataset with 4,435 official training observations and 2,000 official test observations[46] . Each instance contains 36 numeric measurements from a 3 × 3 neighborhood (four bands per pixel) and one central-pixel label in the six observed categories 1, 2, 3, 4, 5, and 7. Table 1 summarizes the fixed experimental input and evaluation boundary established in the notebook.

Table 1. Dataset and official experimental partition

Dataset attribute	**Study-specific protocol**
Public source	UCI Statlog Landsat Satellite (official sat.trn and sat.tst files)
Research endpoint	Six observed cover classes: 1, 2, 3, 4, 5, 7
Analysis unit	Each row: 3 × 3 × 4 = 36 numeric predictors and central-pixel class
Proposed partition	4,435 train; 2,000 held-out test; internal stratified 75:25 validation on sat.trn

Dataset attribute	Study-specific protocol
Processing / quality audit	Pipeline-based feature standardization; no test-set fit or target-based features
Annotation / scope check	No geolocation or scene IDs; spatial-block generalization cannot be assessed

As Table 1 shows, the official sat.trn and sat.tst files remain separate; the 2,000 test records are reserved for final evaluation. The dataset does not provide scene coordinates or spatial blocks, so geographic transferability cannot be estimated. Feature scaling is fitted within the training data by each scikit-learn pipeline, avoiding scaling the entire training-and-test collection together. Table 2 describes the operations actually implemented in the notebook and the boundary separating parameter selection from final testing.

Table 2. Experimental data flow and leakage-prevention controls

Stage	Required handling
A: load	Read official train/test files; inspect class labels and 36-feature dimensions.
B: split	Preserve official test; use stratified train/validation split (75:25) with seed 42.
C: fit	Build raw 36-feature and expanded 46-feature representations; fit scaling inside pipelines.
D: optimize	8 Optuna TPE trials for C; maximize validation macro-F1 on 46 features; refit on official train.
E: evaluate	Test on official sat.tst; export metrics, full reports, confusion matrices, and learning curves.

Table 2 indicates that TPE operates only on a stratified 75:25 split of the official training set. The selected logistic-regression model is subsequently refitted on all 4,435 training cases and evaluated once on the official 2,000-case test set. Probability calibration, independent geographic-block validation, and repeated-seed confidence intervals were not computed.

2.3. Research Framework

Figure 1 compares three actual experimental arms under the same official Statlog train/test partition: (i) standardized logistic regression using the original 36 inputs, (ii) logistic regression on 46 inputs after adding ten quadratic central-band terms, and (iii) the same 46-feature representation with TPE-selected regularization. Keeping the untuned polynomial arm separates the effect of representation from the effect of regularization search.

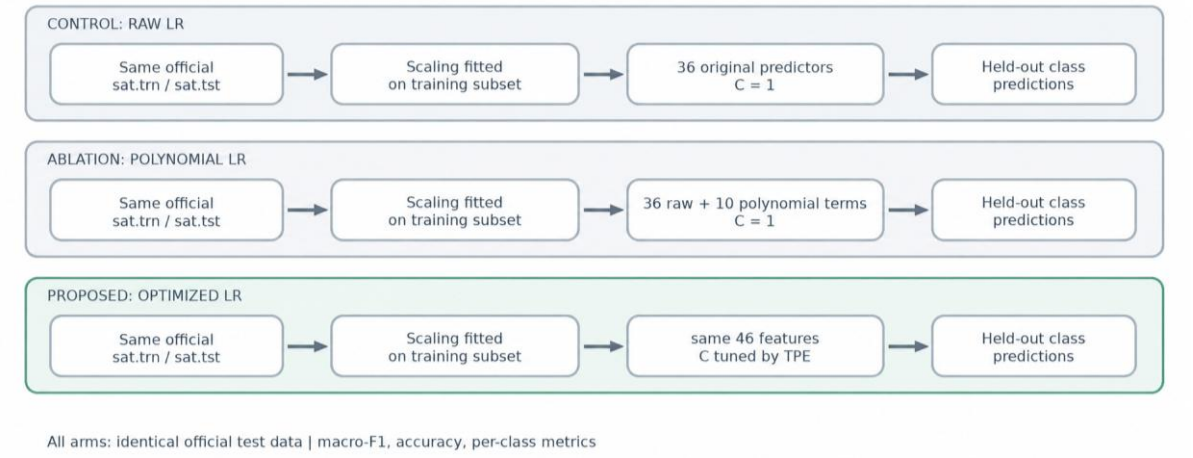


Figure 1. Three-model baseline, feature-expansion, and optimization framework

As shown in Figure 1, the original 36-dimensional record is preserved for every arm. The ten added terms comprise four squares and six pairwise products of the central pixel's four spectral bands. A standard scaler is fitted only on the respective training sample. The test set is never passed to the TPE objective. The matched controls summarized in Table 3 establish which part of the feature-design and optimization sequence explains the observed test-set changes.

Table 3. Controlled baseline, feature expansion, and regularization ablation

Component	Comparative specification
36-feature baseline	Logistic regression; C=1; 36 original numeric inputs.
46-feature untuned	36 raw + 4 center-band squares + 6 pairwise products; C=1.
46-feature optimized	Identical 46 features; Optuna TPE selects C based on training-only validation.

Component	Comparative specification
Controlled factors	Same official train/test set, standardization pipeline, label set, max_iter=1700 and seed 42.
Primary endpoints	Held-out accuracy, macro-F1, weighted-F1, balanced accuracy and per-class precision/recall.

Table 3 defines an empirical ablation rather than a proposed future experiment: all three models are already fitted and evaluated in the supplied notebook. Only the full 36-feature baseline and its 46-feature alternatives were implemented; a separate four-band-only baseline and a degree-optimization experiment were not conducted.

2.4. Proposed Method

The implemented approach is an L2-regularized multinomial logistic-regression classifier with fixed second-degree central-band interactions and TPE selection of the inverse regularization strength C. Figure 2 identifies exactly which operations were executed. In particular, the polynomial degree is fixed at two and the coefficient penalty is the default L2 regularization; sparse L1 penalties or degree searches were not implemented.

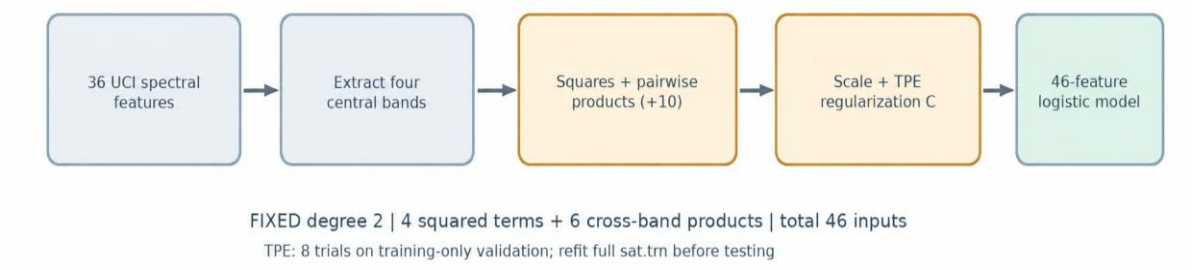


Figure 2. Implemented polynomial feature expansion and TPE regularization pipeline

As illustrated in Figure 2, X is reshaped to (3, 3, 4) only to extract four center-band values. Their four squared values and six unique pairwise products are concatenated to the raw 36-feature vector, producing 46 predictors. The StandardScaler and LogisticRegression(C, max_iter=1700, random_state=42) pipeline is trained after this deterministic feature expansion. Optuna uses a seeded Tree-structured Parzen Estimator over C in [0.0005, 100] on a logarithmic scale, with eight trials in FAST_MODE=True.

Let $x \in \mathbb{R}^{36}$ denote one original patch vector and let $c_b = X_{22b}$ be its central value in band $b \in \{1,2,3,4\}$ after reshaping to $3 \times 3 \times 4$. The four squared central values and six unique cross-band products are collected in (1). Concatenation with the original inputs gives the 46-dimensional representation in (2); no other quadratic terms are generated.

$$\begin{aligned} s_2(c) &= [c_1^2, c_2^2, c_3^2, c_4^2]^T \in \mathbb{R}^4, \\ q_2(c) &= [c_1c_2, c_1c_3, c_1c_4, c_2c_3, c_2c_4, c_3c_4]^T \in \mathbb{R}^6. \end{aligned} \tag{1}$$

$$\phi_2(x) = [x^T, s_2(c)^T, q_2(c)^T]^T \in \mathbb{R}^{46}. \tag{2}$$

For fitting observations $x_1, \dots, x_n$, the mean m_j and population standard deviation s_j of transformed feature j are defined in (3). The fitted scale $\tilde{s}_j$ equals s_j for a nonconstant feature and one for a constant feature. These same fitting-derived quantities transform validation and test observations; they are not re-estimated from those observations.

$$\begin{aligned} m_j &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi_{2,j}(x_i), \quad s_j = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\phi_{2,j}(x_i) - m_j\right)^2}, \\ z_j(x) &= \tfrac{\phi_{2,j}(x) - m_j}{\tilde{s}_j}, \qquad j = 1, \dots, 46. \end{aligned} \tag{3}$$

The class set is $\mathcal{L} = \{1,2,3,4,5,7\}$. Each class k has coefficient vector $w_k \in \mathbb{R}^{46}$ and intercept β_k. The linear scores and multinomial probabilities are given in (4), with W denoting the matrix whose rows are w_k^T.

$$\begin{aligned} a_k(x) &= w_k^T z(x) + \beta_k, \\ p_k(x) &= \tfrac{\exp(a_k(x))}{\sum_{\ell \in \mathcal{L}} \exp(a_\ell(x))}, \qquad k \in \mathcal{L}. \end{aligned} \tag{4}$$

For unweighted fitting records with labels y_i, the L2-regularized multinomial objective can be written in the summed-loss form (5). This form makes the role of the inverse regularization strength $C > 0$ explicit. The intercepts

are not penalized, and a larger C reduces the coefficient penalty relative to the data loss. The predicted class follows (6).

$$\min_{W,\beta}\left[-\sum_{i=1}^{n}\log p_{y_i}(x_i)+\frac{1}{2C}\sum_{k\in\mathcal{L}}\parallel w_k\parallel_2^2\right]. \tag{5}$$

$$\hat{y}(x)=\arg\max_{k\in\mathcal{L}} p_k(x). \tag{6}$$

Table 4 records the exact feature representation, regularization objective, model configurations, and validation boundaries reproduced from the notebook. In contrast to the earlier design-only text, no search over polynomial degree, calibration estimation, or nested cross-validation is claimed.

Table 4. Implemented polynomial logistic-regression and TPE configuration

Design element	Operational definition
Standard input	36-channel numerical patch, with 4 center-band measurements extracted by reshape(3,3,4).
Feature transformation	Fixed degree 2: 4 squares + 6 pairwise interactions; total dimension 46.
TPE optimization	C ~ log-uniform(0.0005, 100), 8 trials, seeded TPESampler(seed=42).
Output model	StandardScaler + L2-regularized multinomial logistic regression; max_iter=1700.
Objective	max_C F1_macro on the 25% stratified internal validation subset.
Ablation	LR 36 features (C=1); LR 46 features (C=1); LR 46 features (C=19.52521).
Scope / limitation	Official 2,000-case test set; no spatial CV, degree search, probability calibration or repeated seeds.

For each class k, let TP_k, FP_k and FN_k denote the true-positive, false-positive and false-negative counts on the internal validation subset. The search maximizes their unweighted six-class macro-F1 in (7).

$$\mathrm{MacroF1}=\frac{1}{6}\sum_{k\in\mathcal{L}}\frac{2\mathrm{TP}_k}{2\mathrm{TP}_k+\mathrm{FP}_k+\mathrm{FN}_k}. \tag{7}$$

Let $\mathcal{H}_8$ contain the eight evaluated values of C, sampled on a logarithmic scale from the stated interval. Equation (8) selects the best evaluated value, where $\hat{y}_{\mathrm{val}}^{(C)}$ is obtained from a candidate pipeline fitted only on the internal fitting subset. The selected setting is then refitted on all official training records.

$$\begin{aligned} C^{*} &= \arg\max_{C\in\mathcal{H}_8}\ \mathrm{MacroF1}\left(y_{\mathrm{val}},\hat{y}_{\mathrm{val}}^{(C)}\right), \\ \mathcal{H}_8 &\subset [5\times10^{-4},100]. \end{aligned} \tag{8}$$

The optimization objective associated with Table 4 is max_C F1_macro(validation; LR(StandardScaler(φ_2(X)); C)), with the 25% stratified validation subset drawn only from sat.trn. The chosen C = 19.525210947020728 achieved an internal validation macro-F1 of 0.82836 and was frozen before final refitting. This validation score is distinct from the 2,000-case test macro-F1 of 0.8232. Implementation and reproducibility. The experiment uses seed 42, an internal stratified train/validation split with test_size=0.25, eight TPE trials, a maximum of 1,700 solver iterations, degree-two feature construction, and fixed official UCI partitions. No inference-runtime benchmarking, probability-calibration experiment, feature stability analysis, or repeated independent-seed evaluation was performed. The claims below are restricted to these actual outputs.

3. Results and Discussion

3.1. Overall Classification Performance

The three logistic-regression configurations were trained with the same 4,435 official training observations and evaluated on the unchanged 2,000-record Statlog test set. Table 5 presents the mandatory baseline–optimization comparison together with an intermediate untuned polynomial model, allowing the contribution of feature expansion to be distinguished from regularization tuning.

Table 5. Official-test performance of baseline, polynomial, and TPE-optimized logistic regression

Model	Accuracy (%)	Macro F1 (%)	Weighted F1 (%)	Balanced acc. (%)
Baseline LR (36 features)	83.95	79.71	82.96	79.38
Polynomial LR, C = 1 (46)	84.20	80.41	83.42	80.06
(Proposed Method) Polynomial LR + TPE (46)	85.30	82.32	84.85	82.04

As Table 5 indicates, the optimized polynomial model obtained 85.30% accuracy and 82.32% macro-F1, compared with 83.95% accuracy and 79.71% macro-F1 for the raw-feature baseline. These are gains of 1.35 and 2.61 percentage points, respectively. Polynomial expansion without tuning yielded 84.20% accuracy and 80.41% macro-F1, whereas subsequent regularization selection added another 1.10 and 1.91 points. This staged pattern supports a measured incremental improvement under the fixed test protocol; it is not evidence of statistical significance across independent runs.

3.2. Per-Class Classification and Confusion-Matrix Analysis

Average scores conceal uneven class behavior. Table 6 therefore reports the complete per-class precision, recall, F1-score, and sample support for the baseline and TPE-optimized polynomial model. The intermediate polynomial classifier remains summarized in Table 5, with its full report available in the supplied notebook.

Table 6. Complete baseline and optimized per-class classification report on the official test set

Class	Support	Base P	Base R	Base F1	TPE P	TPE R	TPE F1
1	461	0.9680	0.9848	0.9763	0.9804	0.9761	0.9783
2	224	0.9498	0.9286	0.9391	0.9636	0.9464	0.9550
3	397	0.8558	0.9270	0.8900	0.8741	0.9270	0.8998
4	211	0.5079	0.3033	0.3798	0.5309	0.4076	0.4611
5	237	0.8404	0.7553	0.7956	0.8522	0.8270	0.8394
7	470	0.7477	0.8638	0.8016	0.7756	0.8383	0.8057
Macro avg	2000	0.8116	0.7938	0.7971	0.8295	0.8204	0.8232
Weighted avg	2000	0.8283	0.8395	0.8296	0.8467	0.8530	0.8485

Table 6 shows that class 4 remained the least reliably recognized category. Its recall increased from 30.33% to 40.76%, and its F1-score from 0.3798 to 0.4611, but these results still indicate a high omission rate. Class 5 recall rose from 75.53% to 82.70%, whereas class 7 recall fell from 86.38% to 83.83%; thus, aggregate improvement was not uniform across classes. The following confusion-matrix figures identify the corresponding error distributions. The count and row-normalized confusion matrices for the baseline are displayed in Figure 3. Values on the diagonal represent correctly classified observations; off-diagonal cells reveal which land-cover categories remain confused.

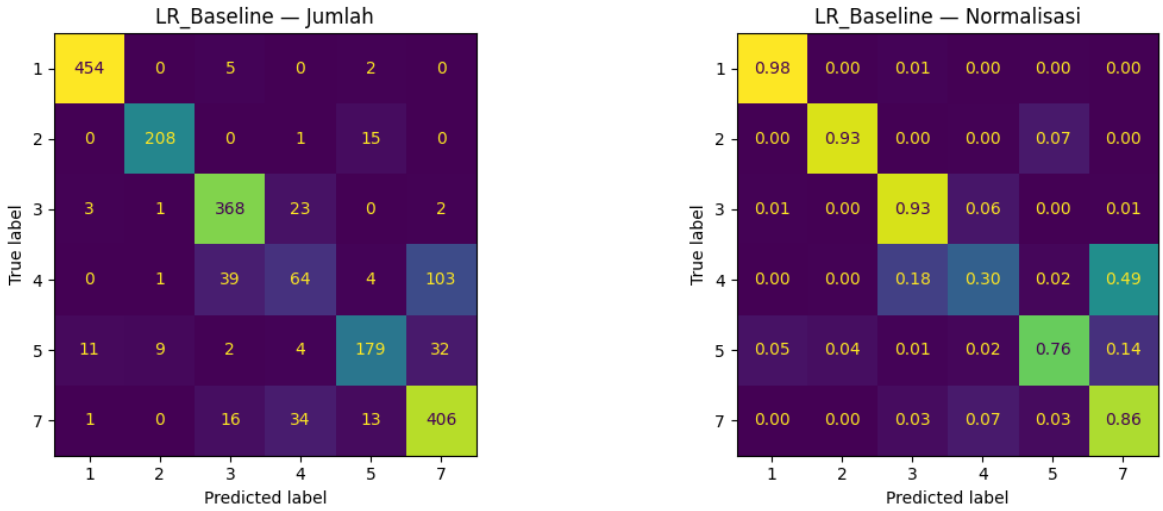


Figure 3. Baseline logistic-regression confusion matrix: counts and row-normalized proportions

Figure 3 illustrates the weak detection of class 4 relative to the other categories: only 64 of its 211 test observations are correctly predicted by the baseline (30.33% recall). This unequal class behavior explains why balanced accuracy and macro-F1 are lower than overall accuracy and motivates examining the optimized model at the class level rather than relying on a single aggregate metric. The corresponding confusion matrices for the optimized 46-feature model are shown in Figure 4, using the same class order and normalization so that differences are directly comparable.

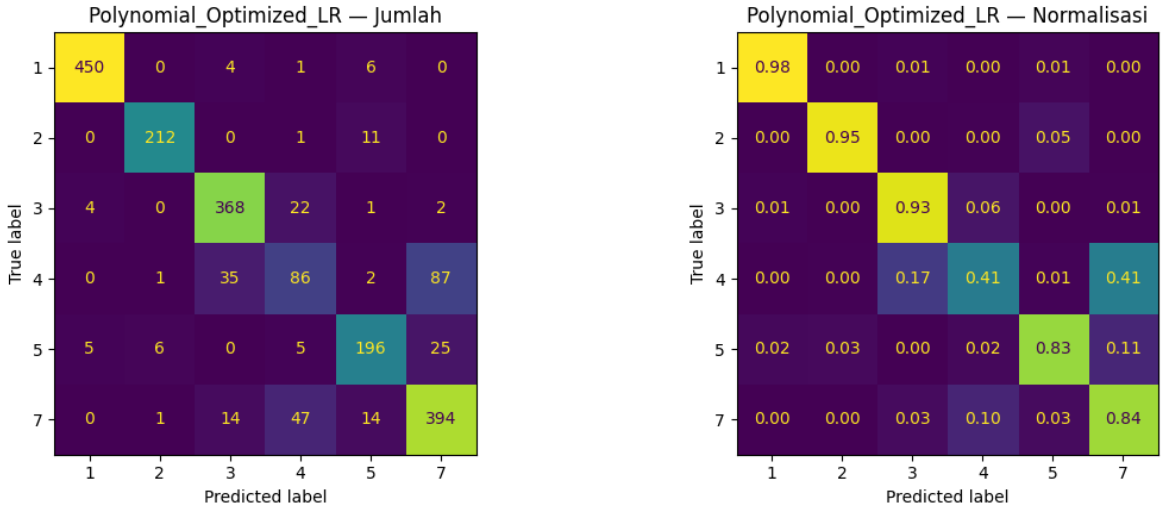


Figure 4. Optimized polynomial logistic-regression confusion matrix: counts and row-normalized proportions

As Figure 4 demonstrates, 86 of the 211 class-4 test samples are now correctly classified, versus 64 for the baseline, although 125 remain misclassified. The improved diagonal is also evident for class 5. Errors involving class 7 persist and its recall declines slightly, confirming that the gain is driven by improvements in selected categories rather than uniform gains over all six classes.

3.3. Learning-Curve Evaluation

Figure 5 compares training and internal-validation macro-F1 as the fraction of the official training data increases. Both curves were computed only from a stratified split of sat.trn; the held-out 2,000 official test observations were not used to construct the learning curves.

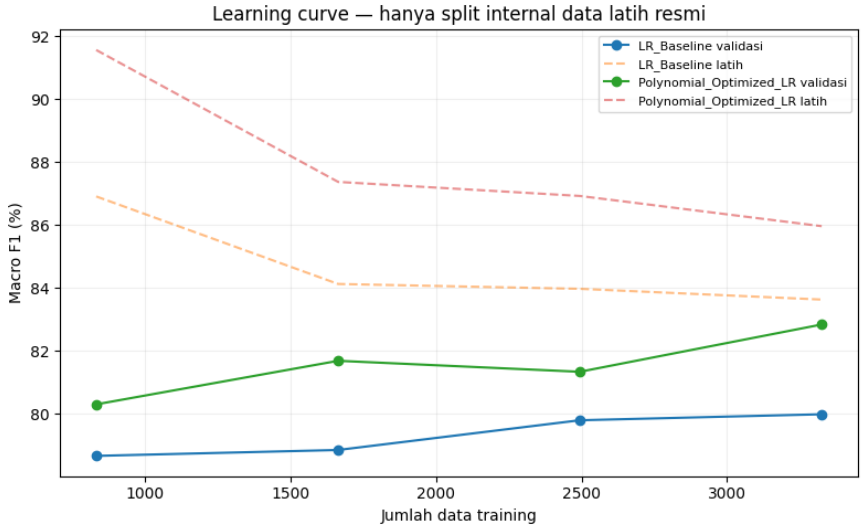


Figure 5. Training and internal-validation macro-F1 learning curves for baseline and optimized logistic regression

Figure 5 provides a diagnostic of sample-size sensitivity rather than an independent test of transferability. Changes in the gap between training and validation scores indicate whether additional training observations help stabilize the fitted classifier, but the plotted points share one validation partition. For this reason, the learning curve cannot establish robust superiority across repeated seeds or geographically distinct scenes.

3.4. Discussion and Limitations

The results indicate that explicit second-order interactions provide a small improvement even with the default regularization strength, while TPE-selected C supplies a larger additional macro-F1 gain. Because the model is linear in its expanded features, this mechanism adds nonlinear dependence on the original center-band values

without introducing a deeper classifier. The selected C of 19.5252 and its internal validation macro-F1 of 0.82836 should be interpreted as configuration-selection information, not an additional test-set performance estimate. Classification errors persist particularly in class 4; improvements in macro-F1 cannot be extrapolated to full-scene maps or unseen acquisition conditions. These boundaries are consistent with current cautions regarding spatial dependence, search-driven optimism, and the need for independently evaluated performance.

The study used FAST_MODE=True with eight TPE trials, a single seed (42), and one stratified validation split, and did not evaluate calibration, log loss, scene-separated validation, or confidence intervals. Hence, the numerical improvements are descriptive for the fixed UCI test set and should not be reported as statistically significant. Repeated or nested cross-validation on the official training data, larger preregistered hyperparameter searches, calibrated-probability diagnostics, and genuinely independent scene labels would be appropriate next steps.

4. Conclusion

This study evaluated whether simple second-order spectral interactions and validation-selected regularization improve multinomial logistic regression on the official Statlog Landsat Satellite benchmark. Ten squared and pairwise center-band terms expanded the original 36 predictors to 46. The optimized model achieved 85.30% accuracy and 82.32% macro-F1 on 2,000 held-out test samples, compared with 83.95% and 79.71% for the baseline, representing gains of 1.35 and 2.61 percentage points. The untuned polynomial model attained 84.20% accuracy and 80.41% macro-F1, distinguishing the effects of feature enrichment and regularization search. Class-4 recall improved from 30.33% to 40.76%, although persistent errors and a slight reduction in class-7 recall indicate uneven benefits. The experiment establishes an observed improvement under one fixed split, not statistical superiority or geographic transferability. Future work should evaluate repeated seeds, more extensive optimization, probability calibration, and scene-independent validation.

References

[1] 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, Dec. 2023, doi: 10.1016/j.isprsjprs.2023.11.014.

[2] 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, Nov. 2023, doi: 10.3390/s23218966.

[3] 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, May 2024, doi: 10.1016/j.jag.2024.103866.

[4] 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, Jan. 2024, doi: 10.3390/rs16030533.

[5] A. F. Jocea, L. Porumb, L. Necula, and D. Raducanu, "Sentinel-2 Land Cover Classification: State-of-the-Art Methods and the Reality of Operational Deployment—A Systematic Review," *Sustainability*, vol. 17, no. 22, p. 10324, Nov. 2025, doi: 10.3390/su172210324.

[6] 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, Jan. 2026, doi: 10.3390/a19010037.

[7] J. Li, S. Xiang, and X. Song, "Screening Nonlinear miRNA Features of Breast Cancer by Using Ensemble Regularized Polynomial Logistic Regression," *Journal of Computational Biology*, vol. 31, no. 7, pp. 670–690, Jul. 2024, doi: 10.1089/cmb.2023.0289.

[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, Nov. 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, May 2025, doi: 10.1109/TNNLS.2024.3413799.

[10] 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, Jul. 2025, doi: 10.1016/j.neunet.2025.107350.

[11] 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, Mar. 2025, doi: 10.1016/j.eswa.2024.125962.

[12] Y. Ma *et al.*, "A Spatial–Spectral Transformer for Hyperspectral Image Classification Based on Global Dependencies of Multi-Scale Features," *Remote Sens. (Basel).*, vol. 16, no. 2, p. 404, Jan. 2024, doi: 10.3390/rs16020404.
[13] H. Ding *et al.*, "TBSSF-Net: three-branch spatial-spectral fusion network for hyperspectral image classification," *Opt. Express*, vol. 33, no. 2, pp. 3466–3500, Jan. 2025, doi: 10.1364/OE.550150.
[14] B. Wang, C. Cao, and D. Kong, "SGFNet: Redundancy-Reduced Spectral–Spatial Fusion Network for Hyperspectral Image Classification," *Entropy*, vol. 27, no. 10, p. 995, Sep. 2025, doi: 10.3390/e27100995.
[15] J. Sun, H. Zhang, X. Ma, R. Wang, H. Sima, and J. Wang, "Spectral–Spatial Adaptive Weighted Fusion and Residual Dense Network for hyperspectral image classification," *The Egyptian Journal of Remote Sensing and Space Sciences*, vol. 28, no. 1, pp. 21–33, Mar. 2025, doi: 10.1016/j.ejrs.2024.11.001.
[16] 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, Nov. 2025, doi: 10.1016/j.optlaseng.2025.109167.
[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, Jul. 2023, doi: 10.1016/j.jag.2023.103364.
[18] J. K. Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," *J. Stat. Softw.*, vol. 106, no. 1, pp. 1–31, Mar. 2023, doi: 10.18637/jss.v106.i01.
[19] 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, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.
[20] S. Friedrich *et al.*, "Regularization approaches in clinical biostatistics: A review of methods and their applications," *Stat. Methods Med. Res.*, vol. 32, no. 2, pp. 425–440, Feb. 2023, doi: 10.1177/09622802221133557.
[21] 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, Mar. 2023, doi: 10.1002/widm.1484.
[22] S. Kapoor *et al.*, "REFORMS: Consensus-based Recommendations for Machine-learning-based Science," *Sci. Adv.*, vol. 10, no. 18, p. eadk3452, May 2024, doi: 10.1126/sciadv.adk3452.
[23] 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, Jun. 2023, doi: 10.1109/TPAMI.2022.3220744.
[24] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.
[25] 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, Sep. 2023, doi: 10.1007/s10994-023-06336-7.
[26] 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, Dec. 2024, doi: 10.1016/j.jag.2024.104288.
[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] C. Wang, R. Shang, Y. Li, J. Feng, and S. Xu, "Cross-scene hyperspectral image classification based on cross-domain feature extraction and category decision collaborative optimization," *Expert Syst. Appl.*, vol. 296, p. 128842, Jan. 2026, doi: 10.1016/j.eswa.2025.128842.
[29] H. Andalib, H. Ebadi, and R. Naanjam, "An intelligent dual-stage fusion framework of optical and radar data for land cover classification," *Remote Sens. Appl.*, vol. 42, p. 101987, Apr. 2026, doi: 10.1016/j.rsase.2026.101987.
[30] W. Zhou, C. Persello, and A. Stein, "Hierarchical building use classification from multiple modalities with a multi-label multimodal transformer network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 132, p. 104038, Aug. 2024, doi: 10.1016/j.jag.2024.104038.
[31] S. He, J. Tian, L. Hao, S. Zhang, and Q. Tian, "Unleashing the full potential of hyperspectral imaging: Decoupled image and frequency-domain spatial–spectral framework," *Expert Syst. Appl.*, vol. 243, p. 122870, Jun. 2024, doi: 10.1016/j.eswa.2023.122870.
[32] C. Shi, J. Chen, and L. Wang, "Hyperspectral image classification based on a novel Lush multi-layer feature fusion bias network," *Expert Syst. Appl.*, vol. 247, p. 123155, Aug. 2024, doi: 10.1016/j.eswa.2024.123155.
[33] Y. Li, X. Yang, D. Tang, and Z. Zhou, "RDTN: Residual Densely Transformer Network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 250, p. 123939, Sep. 2024, doi: 10.1016/j.eswa.2024.123939.

[34] D. Wang, B. Du, and L. Zhang, "Spectral-Spatial Global Graph Reasoning for Hyperspectral Image Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 12924–12937, Sep. 2024, doi: 10.1109/TNNLS.2023.3265560.

[35] X. Zhao *et al.*, "Exploiting foundational spatial semantic prior for accurate few-shot hyperspectral image classification," *Expert Syst. Appl.*, vol. 322, p. 132252, Aug. 2026, doi: 10.1016/j.eswa.2026.132252.

[36] J. Ren, J. Yang, R. Shang, K. Xie, J. Feng, and D. Feng, "Hyperspectral image classification based on multi-scale equivariant feature extraction and geometric equivariant self-attention," *Expert Syst. Appl.*, vol. 300, p. 129909, Mar. 2026, doi: 10.1016/j.eswa.2025.129909.

[37] Z. Zhang, F. Jiang, C. Zhong, and Q. Ma, "HorD 2 CN: High-order deformable differential convolution network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 296, p. 129198, Jan. 2026, doi: 10.1016/j.eswa.2025.129198.

[38] J.-Y. Yang, H.-C. Li, X.-R. Feng, F. Gao, Q. Du, and A. Plaza, "Semisupervised graph U-Net with G-ConvLSTM for hyperspectral image classification," *Expert Syst. Appl.*, vol. 298, p. 129522, Mar. 2026, doi: 10.1016/j.eswa.2025.129522.

[39] M. I. Afjal, Md. N. I. Mondal, and Md. Al Mamun, "Effective hyperspectral image classification based on segmented PCA and 3D-2D CNN leveraging multibranch feature fusion," *J. Spat. Sci.*, vol. 69, no. 3, pp. 821–848, Jul. 2024, doi: 10.1080/14498596.2024.2305119.

[40] 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, Dec. 2024, doi: 10.1080/22797254.2024.2341414.

[41] 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, Nov. 2024, doi: 10.3390/rs16224202.

[42] 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, Jun. 2024, doi: 10.1007/s10596-024-10285-y.

[43] 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, May 2024, doi: 10.1007/s10994-024-06515-0.

[44] 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, Sep. 2025, doi: 10.1016/j.ejrs.2025.06.004.

[45] Y. Chen, Z. Liu, and Z. Chen, "AMS: A hyperspectral image classification method based on SVM and multi-modal attention network," *Knowl. Based. Syst.*, vol. 314, p. 113236, Apr. 2025, doi: 10.1016/j.knosys.2025.113236.

[46] A. Srinivasan, "Statlog (Landsat Satellite)," 1993. doi: 10.24432/C55887.