Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 2, No. 2, [December 2023], pp. [51–62]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Tree-Structured Parzen Estimator–Optimized Support Vector Machines with Spectral–Spatial Feature Fusion for Improved Landsat Land-Cover Classification

Agus Perdana Windarto[1*], Lidya Wati[2], Tengku Mohd Diansyah[3], Erfan Babaee Tırkolaee[4]

[1] *Department of Informatics, Master's Program, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia*
[2] *Faculty of Intelligent Computing, Universiti Malaysia Perlis (UniMAP), Malaysia*
[3] *Computer Science Program, Faculty of Engineering and Computer Science, Universitas Harapan, Medan, Indonesia*
[4] *İstinye Üniversitesi, Istanbul, Turki*

[1] *agus.perdana@amiktunasbangsa.ac.id;* [2] *lidyawati@studentmail.unimap.edu.my;* [3] *dian.10.22@gmail.com;* [4] *erfan.babaee@engr.istinye.edu.tr*
** Corresponding author*

Abstract

Accurate land-cover classification from multispectral satellite data remains challenging because neighboring land-cover classes can exhibit similar spectral responses, while prediction quality depends on both feature representation and model selection. This study assessed the combined effect of spectral–spatial feature fusion and Tree-structured Parzen Estimator (TPE) optimization on a radial-basis-function support vector machine for Landsat land-cover classification. The Statlog Landsat Satellite benchmark provides 4,435 official training instances and 2,000 test instances, each described by 36 spectral measurements from a 3 × 3 neighborhood across six classes. Thirty within-patch spectral–spatial descriptors were appended to the original features, producing 66 predictors, while TPE selected the regularization and kernel parameters from eight trials using a stratified validation split within the training set. A standard support vector machine, an untuned feature-fusion model, and a TPE-optimized feature-fusion model were evaluated on the unchanged official test set. Accuracy increased from 89.55% to 91.15% and 92.45%, respectively; macro-F1 increased from 87.69% to 89.41% and 91.13%. Class-4 recall improved from 57.82% to 65.40%, although confusion among spectrally similar classes persisted. Overall, optimized fusion achieved gains of 2.90 percentage points in accuracy and 3.44 points in macro-F1 over the baseline. These improvements were observed on one fixed benchmark split; repeated-seed evaluation and testing on geographically independent imagery remain necessary to establish robustness and transferability.

Keywords: Landsat land-cover classification; spectral–spatial feature fusion; support vector machine; Tree-structured Parzen Estimator; hyperparameter optimization

1. Introduction

The growing demand for defensible automated decisions in Landsat land-cover classification has made classifying land-cover categories from multispectral satellite measurements an important research problem, because errors can propagate into resource allocation, screening priorities, and operational decisions [1], [2]. The task is not reducible to the maximization of aggregate accuracy, since training-data quality and sampling design influence classification outcomes, while local uncertainty maps reveal reliability differences that a global summary may conceal [3], [4]. Earlier investigators have approached remote-sensing classification problems through conventional discriminative learning, ensemble classifiers, engineered representations, and increasingly sophisticated neural architectures, thereby demonstrating the practical breadth of contemporary machine learning [5], [6]. Across these approaches, representation quality matters because useful class distinctions must be retained without allowing nuisance variation or incidental measurement conditions to dominate the fitted decision boundary][7], [8]. Studies examining model selection also show that hyperparameter search can improve the use of a candidate estimator, yet optimization outcomes depend heavily on search ranges, evaluation folds, and computational budgets [9], [10] . Investigations of structured or spatially dependent observations warn that an apparently excellent random-split score may reflect dependence between training and test samples rather than out-of-sample transferability [11]. Work on

probability reliability further distinguishes correct ranking from reliable decision confidence, making it necessary to consider threshold selection, calibration, and the consequences of overconfident predictions [12]. Application-specific experiments have therefore made valuable progress in feature representation, architecture engineering, and supervised optimization, but each methodological advance addresses only part of the evaluation problem [13], [14]. For satellite land-cover classification, operational evaluation is further complicated by geographic and temporal transfer, spatial autocorrelation, and insufficiently diverse training data, as documented in a recent Sentinel-2 review [15]. Redundancy-reduced and compact spectral–spatial networks offer approaches to controlling representation size and computational complexity, although their neural architectures do not establish the interpretability of an RBF SVM [16], [17]. A comparative satellite-image study evaluated metaheuristic and statistical approaches for SVM tuning, including TPE [18]. For the present experiment, improvements reported on a single split are treated as descriptive differences rather than evidence of statistically stable superiority. The AutoML literature consequently emphasizes defining search spaces, search strategies, and evaluation procedures as connected parts of model development [19]. Hybrid attention–SVM research provides a precedent for combining a learned representation with a margin-based classifier in hyperspectral classification [20]. These considerations motivate the present combination of explicit spectral–spatial descriptors with TPE-optimized support-vector classification. Optimized multisensor models and optical–radar fusion demonstrate other forms of integration, but their different inputs and experimental protocols do not establish the effect of the present descriptor–TPE combination [21], [22]. Cross-domain and cross-scene hyperspectral models address transfer between source and target data [23], [24]. Addressing the present gap instead requires a comparative design with explicit model capacity, partitions, annotation provenance, and preprocessing decisions.

The methodological gap becomes more consequential when TPE-optimized support-vector classification over fused spectral and local spatial descriptors is implemented without jointly specifying the learning objective, search space, resampling strategy, and conditions under which the resulting system will be evaluated [9], [18]. A first technical concern is that flexible tuning procedures may repeatedly exploit validation noise, selecting designs that fit a particular resampling realization while offering no reliable advantage on genuinely independent observations [9] . This concern is especially important in geospatial classification because geographically related observations do not become independent simply through class stratification [11]. Where related measurements originate from one sensor, scene, location, or acquisition period, random partitioning may inflate the apparent gains of both the baseline learner and the optimized alternative [11]. A second concern is that maximizing an undifferentiated metric can neglect low-prevalence outcomes, even when the application places the greatest practical value on detecting the least common or most costly errors [25]. Changing sample weights or introducing synthetic observations may improve empirical class balance, but these operations must occur strictly within the training folds to avoid contaminating held-out evidence [26]. A third issue concerns calibration: a classifier can assign suitable labels while producing probabilities that are unsuitable for risk thresholds, referrals, or resource-limited decision processes [12]. Calibration measures should therefore be assessed alongside discrimination, with threshold choice separated from unbiased model assessment so that a reported improvement reflects a coherent operational objective. Further, the present comparison separates representation refinements from optimizer changes; testing them only as a complete package would not reveal which stage produces an observed gain. For this comparison, changes in feature representation and the number of examined configurations are therefore recorded separately. In resource-constrained settings, additional optimization may also undermine practical applicability if an ostensibly efficient predictive pipeline demands excessive search time or expensive training hardware [10]. Consequently, inference latency and memory consumption are identified as future measurement targets, with hardware to be stated whenever such measurements are reported. The interpretation of Landsat land-cover classification results also depends on the experimental unit: predictions on correlated images or records should not be misrepresented as independent evidence of population-level reliability [11]. External or temporally separated test samples provide a stronger challenge to learned patterns, particularly when nuisance conditions, feature distributions, or target prevalence change after development [11], [15]. Accordingly, this research compares three models on the same official test partition with a prespecified macro-F1 focus, while considering repeated seeds and matched optimization budgets as necessary extensions.

To address these limitations, the present study examines TPE-optimized support-vector classification over fused spectral and local spatial descriptors, with the objective of improving the reliability and practical usefulness of predictions in Landsat land-cover classification under controlled experimental conditions. The proposed contribution is methodological rather than a claim of already demonstrated superiority: it defines a falsifiable comparison of an integrated search-and-learning strategy with matched standard estimators. The optimization procedure should be confined to a prespecified parameter domain, ensuring that each candidate is assessed with an identical

preprocessing pipeline, declared compute budget, and reproducible random seeds [9]. A nested or otherwise leakage-safe validation strategy is appropriate because it separates the selection of tuning settings from the final assessment of predictive benefit [9], [26]. Where the observations exhibit temporal, spatial, scene-level, or acquisition-level dependencies, the allocation of folds must follow those actual dependencies rather than a convenience-driven row shuffle [11]. Where geographic information is available, spatial validation can assess performance beyond a random partition [11]. In the present coordinate-free dataset, macro-F1, balanced accuracy, per-class recall, and confusion analyses characterize the available held-out classification outcomes; class-level assessment complements aggregate accuracy [25]. The model development should also compare the isolated contribution of each named component through matched ablations, so a combined improvement is not erroneously attributed to a single innovation. Appropriate reference systems include a default estimator, a carefully tuned estimator without the proposed addition, and a simplified version that preserves comparable compute or model capacity [9]. To examine stability, the investigation should repeat the optimization and fitting procedures across independent seeds, reporting both central tendencies and variation rather than selecting the most favorable run [9]. Predictive probability quality should be inspected wherever decisions depend on confidence, since performance changes at a fixed threshold need not imply a proportionate improvement in calibrated risk [12]. Computational accountability should include the total search cost and final inference footprint, allowing the research to distinguish an efficient deployed predictor from an expensive model-discovery exercise [9], [10]. If external labels or annotations are incomplete, the evaluation must explicitly bound its conclusions to the validated task and must not substitute patch-level classifications for full-scene geographic mapping evidence. Any data augmentation, feature standardization, imputation, and class-rebalancing operations should be estimated only on the corresponding training partitions before application to held-out instances [26]. The scientific value lies in establishing whether the proposed integration changes the accuracy–reliability–resource trade-off, not merely whether it returns a higher number on one familiar benchmark. By connecting model selection, responsible data partitioning, and component-level validation, this design directly targets the unresolved gap identified above and provides a testable basis for subsequent experiments.

A further distinction is necessary between an optimizer that has already been evaluated for satellite-image SVM tuning and the contribution tested here. A recent comparative journal study evaluated several statistical and metaheuristic optimizers for SVM hyperparameters in satellite imagery and reported favorable results for TPE; therefore, TPE-based SVM optimization alone should not be represented as an unprecedented method [18]. The present experiment instead isolates the effect of the existing 3×3 neighborhood information through the comparison of a 36-feature raw-patch SVM, a 66-feature fused representation and a TPE-tuned fused SVM on identical official test records. The public Statlog records supply four spectral values for each of nine pixels but no georeferenced scene or geographic block identifiers, so neighboring rows cannot be assumed to be neighboring spatial locations. For that reason, the experiment reports class-level classification performance from the predefined UCI partition; it does not assert cross-scene deployment reliability, spatial-block independence, calibrated probabilities or statistically significant superiority from a single run. This explicit scope allows the observed improvement to be attributed to the evaluated configuration while acknowledging the need for repeated trials and scene-independent tests before broader claims can be justified.

2. Method

2.1. Related Work

Recent land-cover research demonstrates the importance of combining spectral bands, vegetation indices, texture and complementary sensor information. Landsat feature-relationship learning, multi-index design and optimized multisensor models offer alternative feature strategies [7], [8], [21]. Optical–radar fusion further illustrates the value of complementary measurements [22]. However, the present Statlog benchmark contains only four-band values from a 3×3 pixel patch, not full-scene or multisensor inputs; local descriptor fusion is therefore evaluated within this narrower scope.

Spectral–spatial methods combine band relationships with local context using optical relation guidance, graph-transformer fusion, multi-scale attention and contextual transformers [13], [14], [27], [28]. Three-branch, redundancy-reduced and dual-branch networks offer different ways to balance complementary information and complexity [16], [29], [30]. Adaptive fusion and compact feature extractors reinforce the value of local interactions for spectrally similar classes [6], [17], [31]. These neural models motivate the representation design but do not provide matched baselines for the smaller Statlog input.

Additional hyperspectral research explores frequency-domain decomposition, hierarchical and transformer fusion, graph pooling and global graph reasoning [32], [33], [34], [35], [36], [37]. Multi-branch convolutional approaches demonstrate another combination of spectral and spatial cues [38] . Hybrid SVM–attention classification suggests that margin-based learners can also benefit from richer representations [20]. Here, explicit statistical descriptors and an RBF SVM are chosen to isolate the effects of feature augmentation and hyperparameter tuning.

Generalization remains difficult when acquisition conditions or target scenes change. Cross-domain mapping, cross-scene optimization, spatial priors, equivariant features, deformable convolutions and semi-supervised graph models address aspects of that challenge [23], [24], [39], [40], [41], [42]. Class-conditional uncertainty analysis shows why aggregate accuracy alone may conceal local errors [4]. Geographic validation can assess spatial optimism if location information exists [11]; the UCI Statlog records do not provide the coordinates required for such evaluation.

A recent satellite-image study has already compared TPE with alternative SVM hyperparameter optimizers, so TPE–SVM alone cannot be presented as a new algorithm [18]. The specific question here is whether descriptors computed from the supplied 3 × 3 patch improve a 36-feature SVM and whether TPE yields an additional gain for the same fused representation. Research on leakage further supports separating final evaluation records from model selection [26].

Automated learning and optimization research emphasizes explicit search spaces, objective functions, computational budgets and held-out comparisons [9], [10], [19], [26]. Calibration matters when model outputs are interpreted as probabilities [12]. Multimodal land-use classification underscores the value of heterogeneous inputs [43], but this study uses only fixed-band neighborhood measurements. Together, these findings motivate a controlled comparison among raw-patch SVM, fused-feature SVM and TPE-tuned fused-feature SVM, without assigning every improvement to tuning alone.

2.2. Dataset

The study used the public UCI Statlog (Landsat Satellite) dataset , comprising 4,435 official training and 2,000 official test records[44] . Each record contains 36 numerical measurements representing four spectral bands over a 3 × 3 pixel neighborhood, with the land-cover class of the central pixel as its label. The six observed classes are 1, 2, 3, 4, 5, and 7. Table 1 summarizes the data and the implemented evaluation boundary; the dataset does not provide geographic coordinates or spatial-block identifiers.

Table 1. Dataset specification and implemented evaluation protocol

Dataset attribute	Study-specific protocol
Public source	https://archive.ics.uci.edu/dataset/146/statlog+landsat+satellite
Research endpoint	six observed Landsat classes; original category code 6 absent
Analysis unit	3×3 patch encoded as 36 spectral inputs and central-pixel label
Proposed partition	Official sat.trn: 4,435 training instances; sat.tst: 2,000 held-out test instances; a 75:25 stratified holdout within sat.trn for TPE selection.
Processing / quality audit	Scaling fitted within each training portion through StandardScaler + RBF SVM pipelines. The notebook did not report checksums, duplicate auditing or imputation.
Annotation / scope check	UCI Statlog already encodes 3×3×4 neighborhoods as 36 features; spatial-geographic coordinates and a reconstructable full scene are not supplied.

Table 1 establishes that the supplied predictors are patch-encoded measurements rather than complete georeferenced imagery. Spectral–spatial descriptors were computed within each record; they were not extracted using adjacency between independent rows. The official UCI test split remained untouched during hyperparameter selection. Consequently, no claim of geographic transfer or spatial-block validation is made. Table 2 describes the actual training, validation and test workflow, including the controls used to prevent model-selection leakage. Additional checks that were proposed but not implemented are addressed as limitations in the discussion.

Table 2. Implemented experimental data flow and leakage-prevention controls

Stage	Required handling
Stage A: audit	Load official UCI sat.trn/sat.tst files; confirm 36 input features and classes 1, 2, 3, 4, 5 and 7. No file hashes were recorded.
Stage B: split	Retain the official 4,435/2,000 train/test split; draw a stratified 75:25 internal holdout from sat.trn (random seed 42).

Stage	Required handling
Stage C: fit	Compute descriptors independently per record; fit StandardScaler on each training subset and evaluate candidate SVMs on internal validation data.
Stage D: freeze	Select C and gamma with eight TPE trials (FAST_MODE=True), then refit the optimized SVM using all 4,435 training records.
Stage E: report	Evaluate three models on the 2,000-instance official test set using accuracy, macro-/weighted-F1, balanced accuracy, classification reports and confusion matrices; generate an internal-holdout learning curve.

As summarized in Table 2, hyperparameter optimization used only sat.trn and its stratified validation subset. Classification reports, confusion matrices and comparison scores were subsequently obtained from the official sat.tst partition. Spatial group validation, probability calibration, feature-stability analysis and repeated optimization seeds were not implemented in this notebook.

2.3. Research Framework

Figure 1 presents the three model variants actually evaluated on the same official training and test partitions: a baseline radial basis function (RBF) SVM with all 36 raw spectral measurements, an untuned SVM with 66 fused features, and a TPE-optimized SVM with the same 66 fused features. This separation enables a direct comparison of feature-fusion and hyperparameter-search effects on identical test instances.

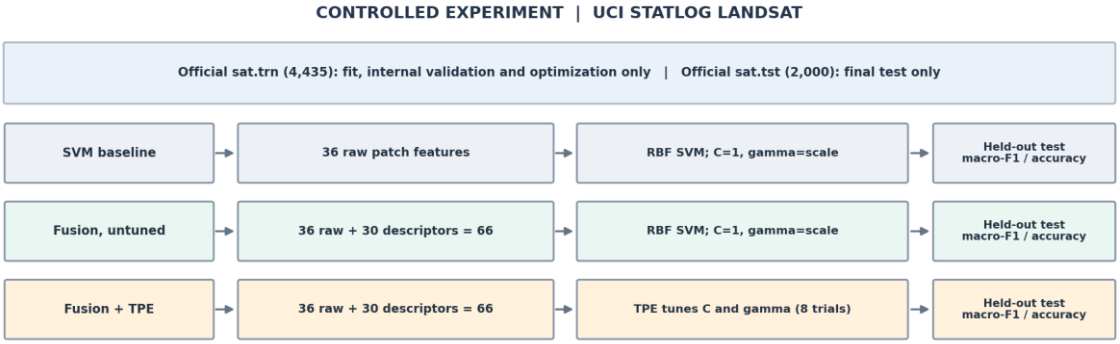


Figure 1. Implemented three-arm baseline and proposed-model comparison

Figure 1 shows that all model variants share a common official UCI test partition, with model fitting and parameter selection confined to the training data. The baseline and untuned fusion arm use C = 1 and gamma = scale; only the proposed fused-feature arm is tuned. Hence, matched test data and default-parameter controls are available, but equal optimization budgets across all arms are not claimed. Table 3 summarizes the controlled comparisons implemented in the notebook. A center-pixel-only model was discussed in the initial study plan but was not run and therefore is not presented as an experimental ablation.

Table 3. Implemented baseline, feature-fusion and optimization comparison

Component	Comparative specification
Baseline	RBF SVM trained on all 36 original spectral patch features (C = 1, gamma = scale).
Proposed innovations	Append 30 spectral–spatial descriptors to the 36 raw predictors; optimize the fused-feature RBF SVM C and gamma with TPE.
Controlled factors	Identical official train/test partitions, input labels, standardized SVM pipelines and final evaluation metrics; seed 42.
Required ablations	Actually tested: 36-feature SVM baseline; 66-feature untuned fusion SVM; 66-feature TPE-optimized fusion SVM. Center-only SVM was not tested.
Primary comparisons	Held-out test-set accuracy, macro-F1, weighted-F1, balanced accuracy and per-class precision, recall and F1-score.

Table 3 distinguishes the gain attributable to descriptor fusion (baseline versus untuned fusion) from the additional gain attributable to TPE search (untuned fusion versus optimized fusion). Because a center-only model and repeated randomized trials were not implemented, conclusions are restricted to the three tested configurations and the reported single run.

2.4. Proposed Method

The proposed method combines within-patch feature engineering with Tree-structured Parzen Estimator (TPE) hyperparameter search for an RBF SVM. Figure 2 presents the implemented feature dimensions and selection procedure. Unlike the initial conceptual diagram, the actual fused representation retains all 36 original spectral inputs instead of replacing them with central-pixel spectra alone.

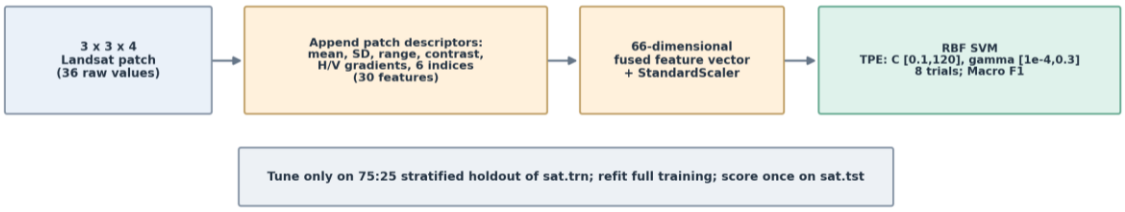


Figure 2. Implemented spectral–spatial descriptor fusion and TPE optimization

As depicted in Figure 2, each original 3 × 3 × 4 patch contributes 36 raw values and 30 appended descriptors: four band means, four standard deviations, four band ranges, four center-versus-neighborhood contrasts, four horizontal and four vertical gradients, and six pairwise normalized central-band indices. StandardScaler and RBF SVM are fitted in a single pipeline. TPE searches C between 0.1 and 120 and gamma between 10^-4 and 0.3 on logarithmic scales.

For one observation, let X_{rcb} denote the value at row r, column c and band b, where $r, c \in \{1,2,3\}$ and $b \in \{1,2,3,4\}$. The central value is X_{22b}. The band mean and population standard deviation are defined in (1). The range and center-versus-surround contrast are given in (2), with $\mathcal{N} = \{(r,c): r,c \in \{1,2,3\}, (r,c) \neq (2,2)\}$ denoting the eight noncentral positions.

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

$$\begin{aligned} R_b &= \max_{r,c} X_{rcb} - \min_{r,c} X_{rcb}, \\ \Delta_b &= X_{22b} - \tfrac{1}{8}\textstyle\sum_{(r,c)\in\mathcal{N}} X_{rcb}\,. \end{aligned} \tag{2}$$

The horizontal and vertical descriptors compare the mean values of opposite patch columns and rows, as in (3). For each pair of central bands, (4) gives a normalized difference index. The absolute values and $\varepsilon = 10^{-8}$ reproduce the denominator used in the implementation, and the condition $1 \le a < b \le 4$ produces six indices.

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

$$I_{ab} = \frac{X_{22a} - X_{22b}}{|X_{22a}| + |X_{22b}| + \varepsilon}, \qquad 1 \le a < b \le 4. \tag{4}$$

Let $x \in \mathbb{R}^{36}$ be the original patch vector. Stacking each four-band descriptor into its corresponding vector and the six indices into I gives the fused representation in (5). Its dimension is $36 + 6(4) + 6 = 66$, so the raw spectral values are retained alongside the 30 added descriptors.

$$\phi(x) = [x^T, \mu^T, \sigma^T, R^T, \Delta^T, (G^h)^T, (G^v)^T, I^T]^T \in \mathbb{R}^{66}. \tag{5}$$

For feature j, let m_j be its mean estimated from the current fitting subset and s_j the corresponding fitted scale. StandardScaler uses the population standard deviation for a nonconstant feature and a scale of one for a constant feature. Each candidate therefore transforms fitting and validation observations using the same training-estimated quantities in (6); the proposed fused pipeline has $j = 1, \ldots, 66$.

$$z_j(x) = \frac{\phi_j(x) - m_j}{s_j}, \qquad j = 1, \ldots, 66. \tag{6}$$

The RBF kernel operates on these standardized vectors according to (7), where $\gamma > 0$ controls kernel width. For each pair of classes, the soft-margin SVM solves the dual problem in (8). Here n is the number of fitting

observations in that class pair, $t_i \in \{-1, +1\}$ is its binary label, and α_i is a dual coefficient bounded by the regularization parameter C. The six-class SVC combines the decisions of its 15 one-versus-one classifiers.

$$K(z_i, z_j) = \exp\left(-\gamma \parallel z_i - z_j \parallel_2^2\right), \qquad \gamma > 0. \tag{7}$$

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j t_i t_j K(z_i, z_j), \\ \text{subject to} \quad & 0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} \alpha_i t_i = 0. \end{aligned} \tag{8}$$

Table 4 specifies the executable configuration used for the present analysis rather than prospective design commitments. The reported results came from FAST_MODE=True, with eight TPE candidate trials and random seed 42.

Table 4. Implemented proposed-model configuration and evaluation rules

Design element	Operational definition
Standard input branch	36 raw UCI spectral inputs; reshape to 3 × 3 × 4 for local descriptor calculation; fold-fitted StandardScaler in each SVM pipeline.
Novel block N1	Append 30 within-patch descriptors to 36 raw inputs (66 features total).
Novel block N2	Optuna TPE search of RBF SVM C [0.1, 120] and gamma [1e-4, 0.3], eight log-scale trials on a single stratified internal holdout.
Output model	RBF SVM refitted on all 4,435 sat.trn records with the selected fused-feature configuration and hyperparameters.
Optimization / decision objective	Maximize macro-F1 on internal 25% validation subset from sat.trn; select C and gamma, then refit model before official sat.tst scoring.
Ablation comparison	36-feature baseline SVM; 66-feature untuned SVM; 66-feature TPE-tuned SVM.
Acceptance endpoints	Official sat.tst accuracy, macro-F1, weighted-F1, balanced accuracy, per-class classification report and confusion matrix; learning curve on sat.trn only.

For the class set $\mathcal{L} = \{1,2,3,4,5,7\}$, validation macro-F1 is the unweighted mean of the six class-specific F1 scores in (9). The quantities TP_k, FP_k and FN_k are the true-positive, false-positive and false-negative counts for class k on the internal validation subset.

$$\mathrm{MacroF1} = \frac{1}{6}\sum_{k\in\mathcal{L}} \frac{2\mathrm{TP}_k}{2\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}. \tag{9}$$

Let $\mathcal{H}_8$ contain the eight evaluated pairs of C and γ. Both parameters are sampled on logarithmic scales within the stated bounds. The selected pair is defined by (10), where $\hat{y}_{\mathrm{val}}^{(C,\gamma)}$ denotes the validation predictions of the candidate training-fitted pipeline. Only the internal validation data enter this selection criterion; the scaler and classifier are refitted on all official training records after selection.

$$\begin{aligned} (C^*, \gamma^*) \quad &= \arg\max_{(C,\gamma)\in\mathcal{H}_8} \mathrm{MacroF1}\left(y_{\mathrm{val}}, \hat{y}_{\mathrm{val}}^{(C,\gamma)}\right), \\ \mathcal{H}_8 \quad &\subset [0.1, 120] \times [10^{-4}, 0.3]. \end{aligned} \tag{10}$$

Table 4 formalizes the objective as maximizing validation macro-F1 over SVM C and gamma, not as optimizing performance on the final test set. The notebook selected C = 7.094831 and gamma = 0.028977 after eight trials (best internal validation macro-F1 = 0.897338). The selected hyperparameters were then used to refit the model on the entire training partition before the official test evaluation.

Implementation and reproducibility. The notebook uses random seed 42, StandardScaler inside scikit-learn pipelines, Optuna TPESampler, an internal stratified 75:25 validation split and the official UCI sat.trn/sat.tst partitions. It additionally produces per-class reports, test-set confusion matrix heatmaps, and a separate learning curve computed solely within sat.trn. No repeated-seed uncertainty intervals, probability calibration, feature-stability statistics, hardware benchmarks or geographic block evaluation were generated; these remain future validation requirements.

3. Results and Discussion

3.1. Overall Classification Performance

The official UCI Statlog evaluation comprised 4,435 training records and 2,000 test records across six classes (1, 2, 3, 4, 5 and 7). Experiments compared a 36-feature baseline RBF SVM, a 66-feature untuned fusion SVM, and the

same fusion model tuned with eight TPE trials. Hyperparameters were selected on a stratified holdout from sat.trn; sat.tst was reserved for the final comparison. Table 5 presents performance on the identical held-out test set.

Table 5. Comparison of baseline, feature fusion and TPE-optimized SVM on sat.tst

Model	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)	Balanced accuracy (%)
SVM baseline (36 features)	89.55	87.69	89.25	87.24
Fusion SVM, untuned (66 features)	91.15	89.41	90.81	88.76
Fusion + TPE SVM (66 features)	92.45	91.13	92.23	90.55

As shown in Table 5, feature fusion alone improved accuracy by 1.60 percentage points and macro-F1 by 1.72 points over the baseline. Adding TPE yielded a further 1.30-point accuracy gain and a 1.72-point macro-F1 gain relative to untuned fusion. Overall, the proposed model exceeded the baseline by 2.90 accuracy points and 3.44 macro-F1 points. These are observed single-run differences, not estimates of statistical significance.

3.2. Per-Class Performance and Confusion-Matrix Analysis

To examine the class-specific effects of optimization, Table 6 reports precision, recall, F1-score and support for both the baseline and optimized SVM on the official test set. Macro and support-weighted averages are provided for completeness; each metric is derived from the classification reports recorded in the notebook.

Table 6. Per-class precision, recall, F1-score and support: baseline versus TPE-optimized SVM

Class	Support	Baseline SVM			Fusion + TPE SVM		
		Precision	Recall	F1-score	Precision	Recall	F1-score
1	461	0.9787	0.9957	0.9871	0.9871	0.9935	0.9903
2	224	0.9646	0.9732	0.9689	0.9733	0.9777	0.9755
3	397	0.8705	0.9647	0.9152	0.8843	0.9622	0.9216
4	211	0.7176	0.5782	0.6404	0.8571	0.6540	0.7419
5	237	0.9273	0.8608	0.8928	0.9399	0.9241	0.9319
7	470	0.8526	0.8617	0.8571	0.8928	0.9213	0.9068
Macro avg	**2000**	**0.8852**	**0.8724**	**0.8769**	**0.9224**	**0.9055**	**0.9113**
Weighted avg	**2000**	**0.8924**	**0.8955**	**0.8925**	**0.9237**	**0.9245**	**0.9223**

As shown in Table 6, class 4 remained the most challenging class, although its precision rose from 0.7176 to 0.8571, recall from 0.5782 to 0.6540 and F1-score from 0.6404 to 0.7419. This corresponds to 122 versus 138 correctly identified class-4 observations out of 211. The optimized model also increased class-5 and class-7 recall to 0.9241 and 0.9213, respectively, whereas recall declined slightly for classes 1 and 3. Across the six classes, macro-F1 increased from 0.8769 to 0.9113, and weighted-F1 rose from 0.8925 to 0.9223. Thus, the overall gain does not imply an improvement in every individual metric.

The confusion matrices in Figures 3 and 4 show both absolute predictions and row-normalized values for the baseline and optimized models, respectively. These plots facilitate inspection of the remaining off-diagonal errors instead of relying on aggregate scores alone.

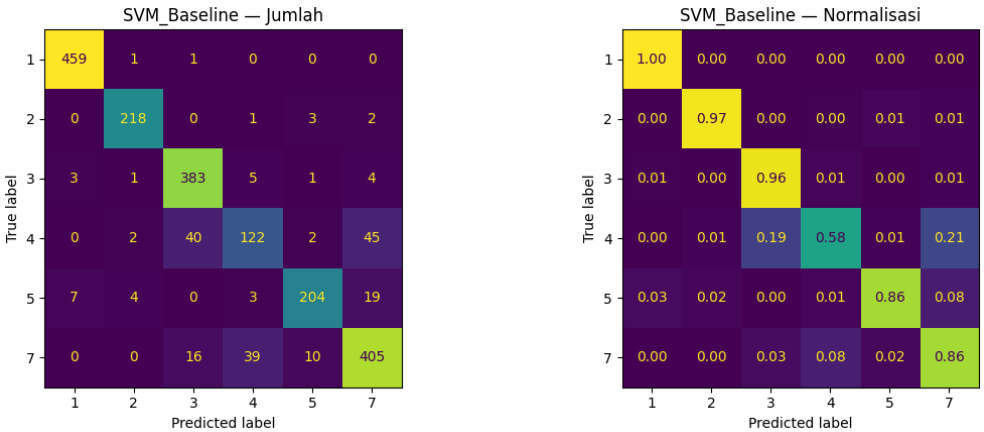


Figure 3. Baseline SVM confusion matrix: absolute counts and row-normalized values

Figure 3 indicates that class 4 presents the largest classification difficulty for the baseline SVM: only 122 of 211 samples (57.82%) were classified correctly, while 40 were assigned to class 3 and 45 to class 7. Moreover, 39 class-7 samples were incorrectly predicted as class 4. These off-diagonal errors reveal substantial overlap among classes 3, 4 and 7 despite the stronger class-wise performance elsewhere, providing a basis for examining the proposed model in Figure 4.

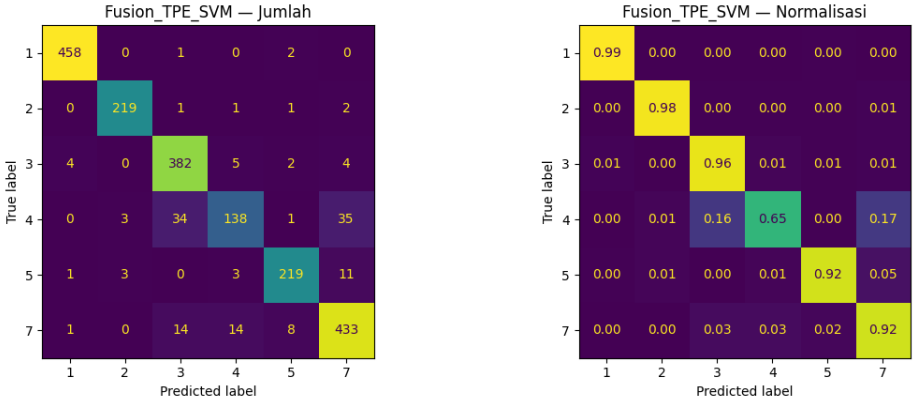


Figure 4. Fusion + TPE SVM confusion matrix: absolute counts and row-normalized values

As shown in Figure 4, class 4 remains frequently confused with classes 3 and 7. In the optimized model, 34 class-4 observations were assigned to class 3 and 35 to class 7. Nevertheless, 138 of 211 class-4 samples were classified correctly, compared with 122 under the baseline in Figure 3. The optimized model also correctly classified 433 class-7 samples compared with 405 for the baseline. These improvements are consistent with the class-level results in Table 6, while the residual errors indicate that spectral overlap has not been fully resolved.

3.3. Learning-Curve Evaluation

Figure 5 presents the learning curves computed using a separate stratified 75:25 split of the official training data only. The curves are therefore development diagnostics, not repeated estimates from the 2,000-instance held-out test partition.

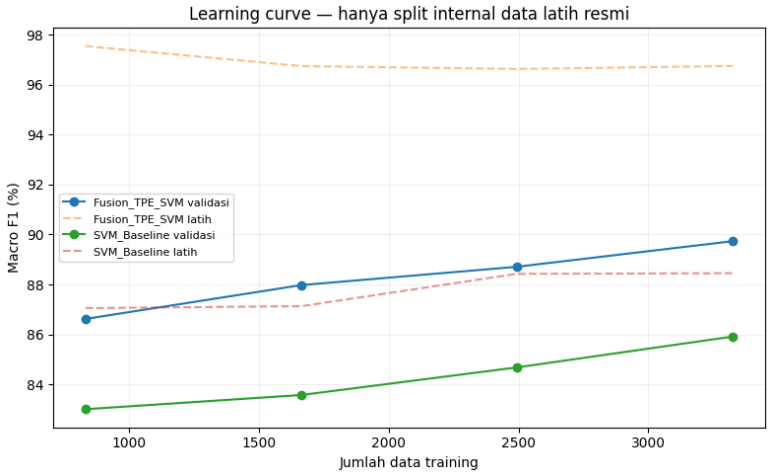


Figure 5. Training and internal-validation macro-F1 learning curves

Learning curves describe how generalization performance varies with training-set size. In Figure 5, the TPE-optimized model shows higher validation macro-F1 than the baseline across the displayed training sizes, with validation performance rising as more training samples are used. Its training score remains markedly above its validation score, indicating a residual generalization gap. Therefore, increased test performance in this benchmark should not be interpreted as proof that overfitting has been eliminated.

3.4. Discussion and Limitations

The findings support a measured improvement from the combination of within-patch spectral–spatial fusion and TPE-selected RBF SVM hyperparameters under the official UCI split. Nevertheless, the FAST_MODE experiment used only eight optimization trials, one seed and one internal validation partition. The dataset contains no

geographic coordinates for true spatial-block evaluation, and the notebook did not evaluate probability calibration, descriptor stability, confidence intervals or inference cost. Multi-seed or nested validation and independent geographically annotated data are needed before asserting broader transferability or statistically robust superiority.

4. Conclusion

This study evaluated spectral–spatial feature fusion and Tree-structured Parzen Estimator (TPE) hyperparameter optimization for support vector machine classification on the official Statlog Landsat Satellite benchmark. The optimized 66-feature model attained 92.45% accuracy and 91.13% macro-F1, improving on the 36-feature baseline (89.55% and 87.69%) by 2.90 and 3.44 percentage points, respectively. Untuned fusion yielded intermediate results (91.15% accuracy and 89.41% macro-F1), supporting complementary benefits from feature enrichment and tuning. Although class-4 recall rose from 57.82% to 65.40%, errors among classes 3, 4, and 7 remained. These findings establish an observed improvement on the fixed test split, not statistical superiority or spatial transferability. Further studies should use repeated or nested validation, broader TPE searches, and geographically independent test scenes to assess stability and generalization.

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, 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, 2023, doi: 10.3390/s23218966.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[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] 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, 2025, doi: 10.3390/su172210324.

[16] 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, 2025, doi: 10.3390/e27100995.

[17] 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.

[18] L. W. Rizkallah, "Optimizing SVM hyperparameters for satellite imagery classification using metaheuristic and statistical techniques," *Int. J. Data Sci. Anal.*, vol. 20, no. 5, pp. 4945–4962, 2025, doi: 10.1007/s41060-025-00762-7.

[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] 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, 2025, doi: 10.1016/j.knosys.2025.113236.

[21] 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.

[22] 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, 2026, doi: 10.1016/j.rsase.2026.101987.

[23] 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.

[24] 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, 2026, doi: 10.1016/j.eswa.2025.128842.

[25] 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.

[26] 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.

[27] 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.

[28] M. Zhang, Z. Lei, L. Liu, K. Ma, R. Shang, and L. Jiao, "Efficient evolutionary multi-scale spectral-spatial attention fusion network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 262, p. 125672, 2025, doi: 10.1016/j.eswa.2024.125672.

[29] 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, 2025, doi: 10.1364/OE.550150.

[30] S. Xu, R. Liu, B. Jia, Z. Xie, and H. Gao, "DBSSFNet: Dual branch spatial-spectral-fusion-net for hyperspectral image classification," *Opt. Laser Technol.*, vol. 194, p. 114473, 2026, doi: 10.1016/j.optlastec.2025.114473.

[31] 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, 2025, doi: 10.1016/j.ejrs.2024.11.001.

[32] 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, 2024, doi: 10.3390/rs16020404.

[33] 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, 2024, doi: 10.1109/TNNLS.2023.3265560.

[34] 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, 2024, doi: 10.1016/j.eswa.2023.122870.

[35] 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, 2024, doi: 10.1016/j.eswa.2024.123155.

[36] 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, 2024, doi: 10.1016/j.eswa.2024.123939.

[37] K. Sarpong *et al.*, "Hyperspectral image classification using Second-Order Pooling with Graph Residual Unit Network," *Expert Syst. Appl.*, vol. 238, p. 122202, 2024, doi: 10.1016/j.eswa.2023.122202.
[38] 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, 2024, doi: 10.1080/14498596.2024.2305119.
[39] X. Zhao *et al.*, "Exploiting foundational spatial semantic prior for accurate few-shot hyperspectral image classification," *Expert Syst. Appl.*, vol. 322, p. 132252, 2026, doi: 10.1016/j.eswa.2026.132252.
[40] 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, 2026, doi: 10.1016/j.eswa.2025.129909.
[41] Z. Zhang, F. Jiang, C. Zhong, and Q. Ma, "HorD2CN: High-order deformable differential convolution network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 296, p. 129198, 2026, doi: 10.1016/j.eswa.2025.129198.
[42] 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, 2026, doi: 10.1016/j.eswa.2025.129522.
[43] 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, 2024, doi: 10.1016/j.jag.2024.104038.
[44] A. Srinivasan, "Statlog (Landsat Satellite)," 1993. doi: 10.24432/C55887.