Land Use/Land Cover Classification Using CatBoost and Multi-Source Remote Sensing Data in the Kurdistan Region of Iraq

Ragheb Kamal Mohammad[1]* and Ameer Abdulazeez Abdullah[1]

[1]Department of Forestry, College of Agricultural Engineering Sciences, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq

*Corresponding author's email: ragheb.mohammad@su.edu.krd

Abstract

Despite recent advances in remote sensing, classifying land use and land cover (LULC) remains difficult, especially in semi-arid areas, because spectral confusion between classes reduces mapping accuracy. In this paper, we present a multi-stage, multi-source feature-fusion approach for six LULC classes (water, built-up, agriculture, dense forest, open forest, and bareland) using the CatBoost classifier in the intersecting parts of the Amedi, Mergasur, and Akre districts of the Kurdistan Region of Iraq (KRI). Features are ordered into seven different scenarios (from A to G), starting with the visible bands (Scenario A) as baseline and progressing up to adding Sentinel-2 NIR/SWIR bands, land surface temperature (LST), spectral indices (NDVI, NDWI, NDBI, BSI), GLCM textures, Sentinel-1 SAR backscatters, and SRTM topographic predictors (DEM, slope, aspect). We used 10-fold cross-validation repeated 3 times (30 runs per scenario; cross-validation set: 1,200 points), and independently validated the best-performing scenario on 600 held-out test points (100 per class). We compared scenarios using Friedman tests, post-hoc Wilcoxon tests (with Holm correction), and SHAP-based interpretability tests. Overall accuracy (OA) of cross-validation improved from 83.39% (Kappa = 0.801, F1 = 83.26%) in Scenario A to 95.50% (Kappa = 0.946, F1 = 95.48%) in Scenario G, which represents an improvement of 12.11 percentage points ($\chi^2F = 131.11$, $p < 0.001$, $n = 30$, $k = 7$). After applying Holm correction, four out of six scenario-pair transitions are statistically significant. The largest improvements came from adding NIR/SWIR bands (+6.14 percentage points) and topographic predictors (+2.31 percentage points). The SHAP and CatBoost feature importance rankings are mostly consistent, ranking slope and elevation as the most important predictors. However, SHAP values reflect a feature's importance in making an individual prediction, not in improving model accuracy. Across 600 independent test points, Scenario G achieved an overall accuracy of 95.33%, a Kappa of 0.944, and a weighted F1-score of 95.35%. These findings demonstrate that integrating multi-source satellite data with an appropriate machine learning method significantly improves LULC classification in semi-arid environments characterised by complex topography.

Keywords: CatBoost classifier, LULC map, multi-feature data, SHAP-based, Sentinel-1/2, DEM, KRI

1. Introduction

Classifying Land Use and Land Cover (LULC) is becoming more crucial in remote sensing. LULC classification has been used to manage the world's natural resources, evaluate the impacts of global environmental changes, and formulate policies to reduce climate risks and enhance sustainability [6, 47]. LULC classification is especially important but difficult in the uneven topography of the post-conflict and ecologically transitional areas such as the Kurdistan Region of Iraq. Here, the rapid loss of forests, the expansion of urban areas, and extreme environmental stresses highlight the need for sustained and highly dynamic spatial assessment [1, 26]. The unmatched detail of the high-resolution, freely available Copernicus Sentinel-1 radar and Sentinel-2

ISSN 2072-3857

multispectral datasets has changed the ability of landscape monitoring and shifted the paradigm to assess previously unapproachable areas. These datasets are revolutionary for monitoring the changes in time and space of critical regions [7, 42]. One of the main difficulties in LULC classification, particularly in semi-arid and rugged areas, is spectral confusion. Spectral confusion occurs when pixels from similar classes, such as bareland and open forest, are frequently confused with each other when using a single source of information [5, 34].

Multi-sensor platforms can capture many different aspects of the natural world at a given location. For example, optical sensors capture elements of vegetation biochemistry and structure, while radar can get vegetation structure and the underlying surface. Thermal sensors obtain the system's energy and moisture balance, and topographic data catch a region's micro-climatology [4].

Random Forest is an initial candidate for this task. Its bagging design creates independent trees and averages their predictions, which is different from boosting techniques in that there is no attempt to minimize residual error across trees [39]. Gradient boosted decision trees (GBDTs) progressively fit trees to correct the errors of the previous trees; XGBoost, a popular implementation of GBDTs, is known to suffer from target leakage and prediction shift on certain tabular datasets, which motivated the creation of CatBoost's ordered boosting framework [32]. Empirical studies have shown mixed results across tabular and remote sensing datasets. Some have argued that, at the broadest level of magnitude, CatBoost performs better than XGBoost and Random Forest [9, 32]. With the increasing role of machine learning in land management, if trust in science is to be built, it will be necessary to explain predictions, including the use of SHAP [19, 20].

Present multi-sensor land cover mapping systems have concentrated on combining optical, radar, thermal, and digital elevation data to aid in classification of complex geographic areas [4, 5, 38]. In the Kurdistan Region particularly, existing research has used LULC classification based on Sentinel-2 through Google Earth Engine [12]. Many studies rely on conventional importance rankings of features to determine important spatial predictors [9, 29, 39]. A gap in the literature is that most studies evaluate feature inputs for each layer in a stack, but have not studied in detail whether adding different aspects of the physical domain improves classification accuracy [5, 7]. Therefore, it is not known how different domain groups, such as SAR backscatter, surface thermal dynamics, or spatial textures, affect classification performance in complex topography [5, 38].

The present study suggests a progressive multi-source feature combination method based on the CatBoost classifier, unlike most existing multi-source LULC research, which treats feature stacks as single, combined entities and fails to evaluate each sensor domain's individual contribution. The method examines classification efficiency across seven successive feature stacks and six LULC classes, cross-validates SHAP feature importances against the actual improvement in classification accuracy for each successive stack, and compares SHAP feature-importance rankings with the observed accuracy increases for each successive stack.

2. Materials and methods

2.1 Study area description

Figure 1 shows the study area, a 400 km² area where the three districts of Amedi, Mergasur, and Akre meet. This study area is situated in the mountainous region of northern Kurdistan Region, Iraq. The zone is located within the Duhok and Erbil Governorates. The approximate location of this study area is around 43.800°E and 37.000°N. The square shape of the area is defined based on the Universal Transverse Mercator (UTM) projection (Zone 38N, WGS 84 datum), with easting coordinates varying from 383,200 m to 403,200 m and northing coordinates varying from 4,085,600 m to 4,105,600 m

[29]. The area is characterized by rough high-relief topography that resulted from the Zagros fold-thrust belt, including steep topographical variation, seasonally flowing drainages, and different surface morphology. The land cover within the local area consists of dense and sparse oak forests, rocky surfaces, bare lands, small settlements, and seasonally flowing water bodies [18, 33].

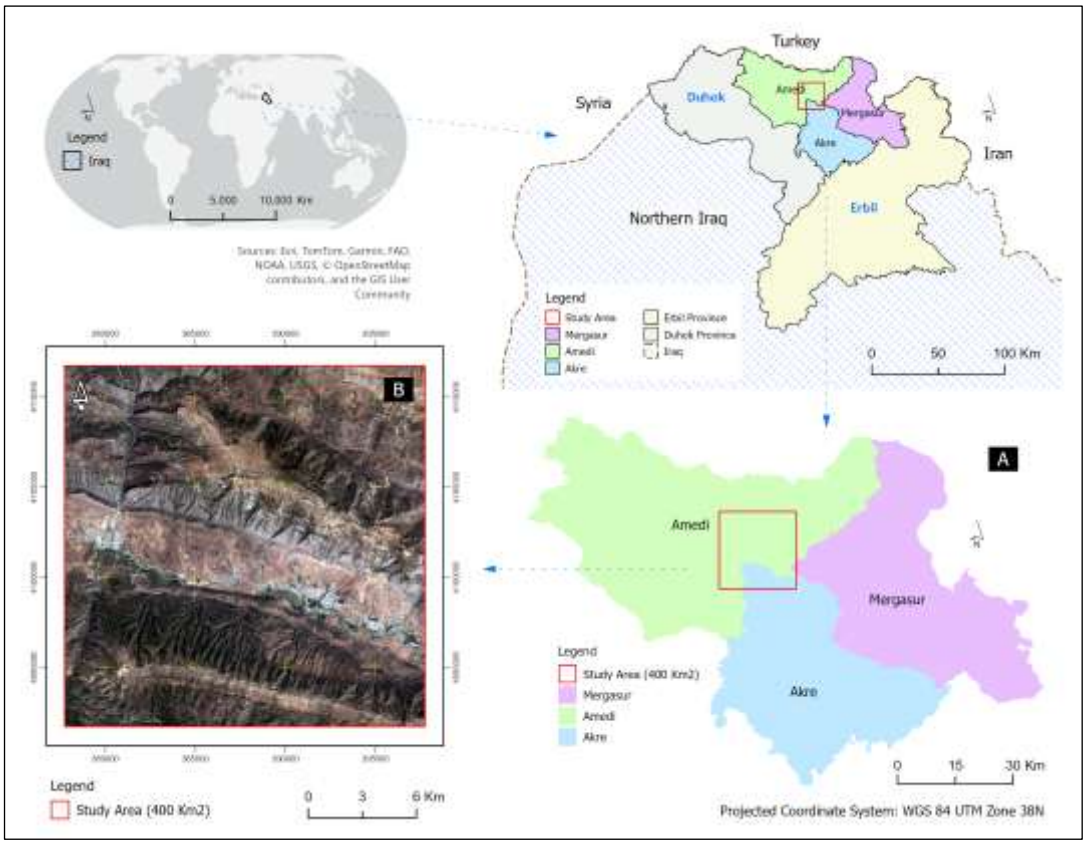


Figure 1. Map of the study area within the Kurdistan Region of Iraq

2.2 Dataset preparation and software application

The inputs from all data sources, Sentinel-2 (collected as a cloud-free temporal median composite), Sentinel-1 SAR, thermal LST (based on Landsat 8/9), and SRTM topography were collected and preprocessed for the period of January–December 2025 in the Google Earth Engine (GEE) platform [13]. To minimize the influence of atmosphere and clouds, the temporal median composites from cloud-free images (with cloud cover <10%) were created. Blue (B2), green (B3), and red (B4) visible bands were used with the resolution of 10 meters. Near infrared (B8) and short-wave infrared (B11, B12) bands, initially 20 meters, were resampled to 10 meters by applying bilinear interpolation in order to preserve subpixel boundaries. This resampling harmonizes the output grid resolution with the native 10 m bands but does not increase the native spatial detail of the original 20 m acquisitions; bilinear interpolation was used to align pixel grids and smooth edge transitions between adjacent 20 m cells, rather than to recover sub-20 m information. The processing of SAR and topographic information included the radiometric calibration, terrain correction, and slope and aspect calculations before stacking [40]. The spectral indices and the texture metrics using the GLCM algorithm were computed in a moving window fashion and matched to other feature layers for the pixel-level correlation. In the final step, all the features were resampled and georeferenced to 10 m and UTM Zone 38N for reproducible inputs in all seven scenarios.

The main pipeline consisted of three software tools dedicated to specific steps: Google Earth Engine, ArcGIS Pro, Google Colab (Python 3), and a Garmin GPS device was used to collect independent field validation points.

In the first place, the preprocessing was carried out in GEE with the following actions: cloud masking, temporal compositing, SAR terrain correction, spectral indices calculation, GLCM texture metrics computation, and topographic information extraction. The result was a 23-band raster-stacked image with the resolution of 10 meters and exported to Google Drive. Then, this stacked raster was imported to ArcGIS Pro, where the Extract Multi Values to Points tool sampled the 23 raster values at 1,200 training points and 600 validation points; each point had its own class value and name, as well as X/Y coordinates, resulting in the creation of the point-attribute table, which matched ground reference points to their feature vectors.

After this procedure, the point-attribute table was exported in CSV format and imported to Google Colab, a cloud-based Python 3 Jupyter notebook interface. Further analyses included CatBoost model training in seven cumulative feature scenarios, 10-fold cross-validation (repeated 3 times), non-parametric statistical testing (Friedman, Wilcoxon, Kendall's W, and McNemar), SHAP interpretation of CatBoost models, independent accuracy assessment on 600 validation points, and LULC classification map generation. The main Python library used in this step was CatBoost [32].

2.2.1 Spectral satellite indices

Five spectral bands; Blue (B2), Green (B3), Red (B4), Near-Infrared (B8), Shortwave Infrared-1 (B11) were derived from the Sentinel-2 reflectance used for the following indices:

NDVI

Derived to separate canopy density and chlorophyll absorption patterns [41].

$$NDVI = \frac{Band8 - Band4}{Band8 + Band4} \ldots (1)$$

NDWI

Extracted to delineate surface moisture and open water boundaries [23].

$$NDWI = \frac{Band3 - Band8}{Band3 + Band8} \ldots (2)$$

NDBI

Derived to identify highly reflective artificial impervious surfaces [46].

$$NDBI = \frac{Band11 - Band8}{Band11 + Band8} \ldots (3)$$

BSI

Calculated to differentiate exposed soils, rocky outcrops, and agricultural fallows from vegetative areas [36].

$$BSI = \frac{(Band11 + Band4) - (Band8 + Band2)}{(Band11 + Band4) + (Band8 + Band2)} \ldots (4)$$

LST

In order to incorporate the microclimate characteristics in the region, the Land Surface Temperature (LST) was extracted directly from the Landsat 8/9 Collection 2 Level-2 Surface Temperature band (ST_B10) through the Google Earth Engine Platform. This pre-computed data by USGS uses the single-channel algorithm where the surface emissivity is calculated from the ASTER Global Emissivity Database (GED) and is corrected for the atmosphere [21].

The values of digital numbers in the band were then converted to Kelvin units based on the USGS-provided scale factor and offset:

$$LST_K = ST_{B10} \times 0.00341802 + 149.0 \ldots (5)$$

and afterward converted to Celsius:

$$LST_{°C} = LST_K - 273.15 \ldots (6)$$

For proper interpretation and mapping purposes, the pixels with any cloud contamination were removed using the QA_PIXEL band before compositing. Further, the thermal map was then re-sampled from the 30 m native resolution to the common 10 m resolution, which will match

the resolution of optical and radar bands. The re-sampling process brings all feature layers to the same grid resolution output but does not add more detail in the native (30 m) thermal band of the Landsat 8/9 imagery and results in a 10 m LST layer that needs to be considered up-sampled [31].

2.2.2 Spatial texture (GLCM) features

To acquire the spatial textures of local areas, Gray-Level Co-occurrence Matrices (GLCM) were calculated by using the GLCM Texture function of Google Earth Engine, which uses 3 as the size of the matrix with a window size of 7 × 7 pixels (size of window side = 2 × size + 1). Calculation of GLCM was done considering four directions for offsets (0°, 45°, 90°, and 135°) at the level of 1 pixel distance. Prior to GLCM calculation, input bands were converted into 8-bit unsigned integer type (256 gray levels). For the Sentinel-2 NIR band (B8), the surface reflectance band was converted from 10,000 scale factor to the 0–255 range, while the Sentinel-1 VV band was linearly scaled (VV × 10 + 200) prior to quantization. From the Sentinel-2 NIR band, four texture descriptors contrast, entropy, homogeneity, and variance were derived, while contrast and entropy descriptors were derived from the Sentinel-1 VV band [45].

2.2.3 Topography Features

DEM

The variables of elevation, slope, and aspect associated with topography were created by reclassifying the Digital Elevation Model (DEM) produced by the Shuttle Radar Topography Mission (SRTM). Resampling to a common 10-meter grid was performed in order to incorporate these variables with other feature families such as optical, radar, textural, and thermal variables. The purpose of this resampling was to match the DEM variables (elevation, slope, aspect) to the common 10 m grid without altering the resolution (30 m) of the original SRTM data [8].

Slope

Slope gradient, in degrees, has been derived from the DEM by calculating the maximum elevation difference between each cell and its neighboring cells through a commonly used third-order difference technique.

$$Slope = \arctan\left(\sqrt{\left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2}\right) \dots (7)$$

where:

$\partial z/\partial x$ and $\partial z/\partial y$ represent the rate of elevation change in the easting and northing directions, respectively [28].

Aspect

Aspect, expressed in degrees (0–360°) and indicating the compass direction of the steepest downslope gradient, was similarly derived from the DEM in Google Earth Engine Platform [14].

2.3 Definition of LULC Classes and Design of Reference Samples

The displayed information in Figure 2 includes six land use and land cover (LULC) classes chosen specifically in order to account for the main types of land cover of the mountainous and semi-arid environment of the study site. These classes include water, built-up areas, agriculture, dense forest, open forest, and bare land.

The current study used a total of 1,800 points, which includes their coordinates X-Y. The first part has 1,200 points, which are stratified training points, evenly distributed in the study area, representing all the six LULC classes with 200 points per class. These 1,200 points were selected entirely through visual interpretation of high-resolution GEE basemap imagery, without field verification. These 1,200 points were used in two separate procedures. First, model assessment and scenario benchmarking (Section 3.1, Table 2) were done using a repeated 10-fold cross-validation technique on the full set of 1,200 points (done three times, resulting in 30 runs per scenario); for each fold, 90% of the samples were used for training (1,080 samples), while 10% were left for validation (120 samples).

Second, a single stratified 70/30 split of the same 1,200 points (840 training / 360 test) was used separately to fit one model per scenario for the feature-importance analysis (Section 3.3, Table 4) and to generate the held-out per-class accuracy assessment and confusion matrix reported in Section 3.4 (Table 5, Figure 6). All the 1,200 points were visually categorized based on the high-resolution basemap images from Google Earth Engine (GEE).

The third part includes 600 points (with X-Y coordinates, 100 points per class), which are independently selected using stratified sampling for final validation of the LULC map. All 600 points had not been included in the process of modeling. Of these 600 points, 300 were collected through field verification and 300 through visual interpretation of high-resolution GEE basemap imagery, giving equal weight (50% each) to field-based ground truth and image-based reference data for defining the independent validation set. Field verification was prioritized for points accessible from the road network across all six classes, while the 300 image-interpreted points covered locations inaccessible for field visits such as steep valleys, high-elevation mountain slopes and center of River that occurred within every class rather than being concentrated in any single one. At least 300 m distance should be kept between all of the training points and all of the validation points in order to prevent spatial dependence. Such distance significantly exceeds both 10 m pixel size of Sentinel-1/Sentinel-2 feature layers and 7×7 pixels (70 m) GLCM texture window (see section 2.2.2) thus preventing any training and validation points to lie on the same feature pixel or its neighbor.

It is worth noting that open forest and dense forest were treated as separate classes not combined into one because of canopy density being important for forest degradation detection in the region in question, as seen from preliminary spectral analysis results[2].

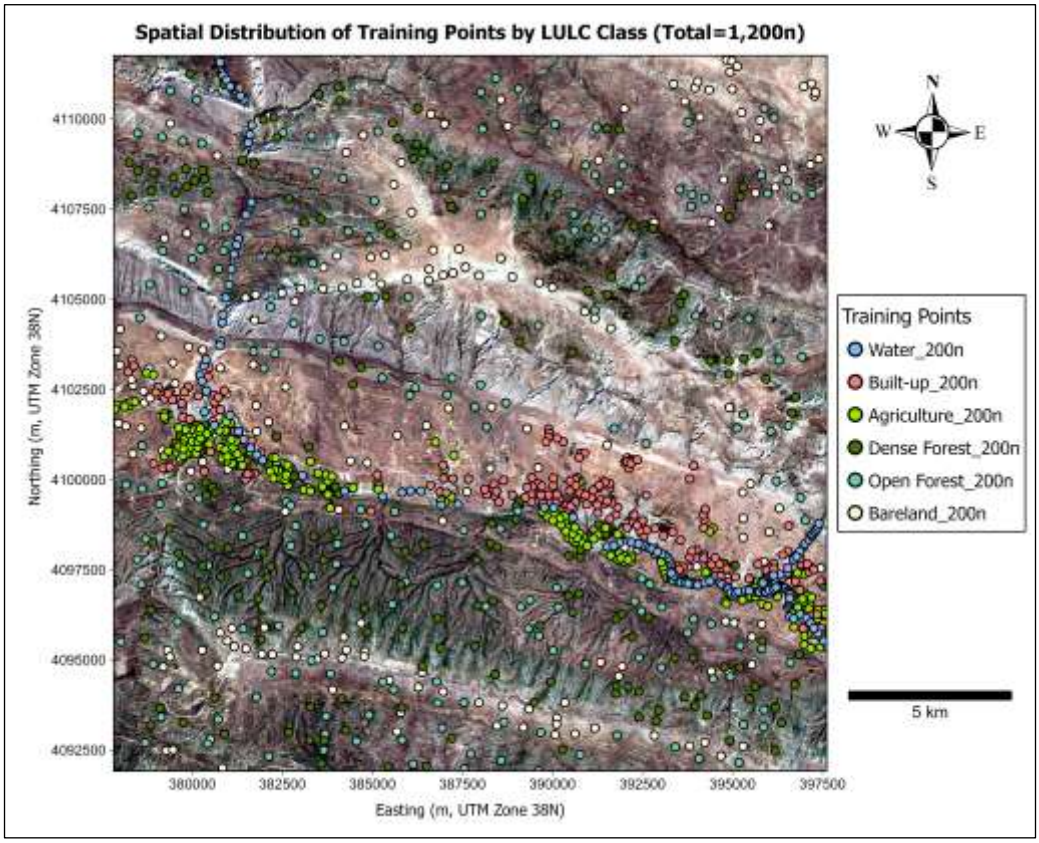


Figure 2. Spatial distribution of the 1,200 training points across the six LULC classes.

2.4 Progressive multi-feature scenarios

Table 1 shows the experimental design, seven cumulative scenarios (A–G) were constructed to assess the incremental accuracy improvement obtained by sequentially adding feature groups to the previous scenario. Because each scenario retains the features of the preceding scenario, the accuracy gain between adjacent scenarios reflects the conditional marginal contribution of the newly added feature group rather than the independent contribution of that data source.

Table 1. Definition of the seven progressive cumulative feature scenarios and their targeted physical domains.

Scenario	Feature	Stack Composition/Sensor Domains	Targeted Physical
A (Baseline)	3	Visible bands (Blue, Green, Red)	True-color baseline spectral signatures
B	6	**Scenario A** + Near-infrared & SWIR bands (B8, B11, B12)	Vegetation canopy & soil moisture dynamics
C	7	**Scenario B** + Land Surface Temperature (LST)	Microclimatic thermal characteristics
D	11	**Scenario C** + Spectral Indices (NDVI, NDWI, NDBI, BSI)	Normalized biochemical and physical boundaries
E	17	**Scenario D** + Optical GLCM Textures (Contrast, Entropy, Homogeneity, Variance) + SAR Textures (Contrast, Entropy)	Localized spatial structure and image texture
F	20	**Scenario E** + SAR Backscatter (VV, VH, VV/VH ratio)	Cloud-free structural geometry and roughness
G	23	**Scenario F** + Topographic Predictors (DEM, Slope, Aspect)	Terrain elevation, incline, and solar exposure limits

2.5 CatBoost classifier

2.5.1 CatBoost algorithm selection and architecture

The classification task was carried out using CatBoost, which is a novel gradient-boosted decision tree (GBDT) method. Traditionally, GBDT models struggle with prediction and target leakage effects in the training phase as these models apply sequential data [32].

However, CatBoost solves the problems by the following means:

(1) Ordered Boosting: a method that uses random permutations of data to calculate the gradients over smaller subsets of data and avoid overfitting.

(2) Symmetric trees, i.e., binary decision trees with similar splitting rules throughout the whole tree level, making them faster and more general [32].

2.5.2 Cross-validation and standard hyperparameter implementation

For a fair controlled baseline comparison for all seven scenarios of cumulative features and for avoiding any bias resulting from hyperparameter tuning, the CatBoost classifier was used with its default hyperparameters: 1000 iterations, tree depth of 6, L2 leaf regularization coefficient of 3.0, and automatically defined learning rate by CatBoost depending on dataset size and iteration number. No changes have been made to the default hyperparameter values. Evaluation was done using strict 10-fold cross-validation performed three times (resulted in 30 runs per scenario) using a training set of 1200 items (200 items per class) [15].

The same stratified folds were preserved for all seven cumulative feature scenarios (using standard-fixed random seed = 42), which

means that each fold (i) includes the same holdout items in every scenario stack. It is a necessary condition for the paired Wilcoxon signed-rank test. At this point, all of the models were evaluated on the independent validation set of 600 points (100 points per class; see Section 2.3) in order to provide an unbiased estimate of regional mapping accuracy [30] (Figure 3).

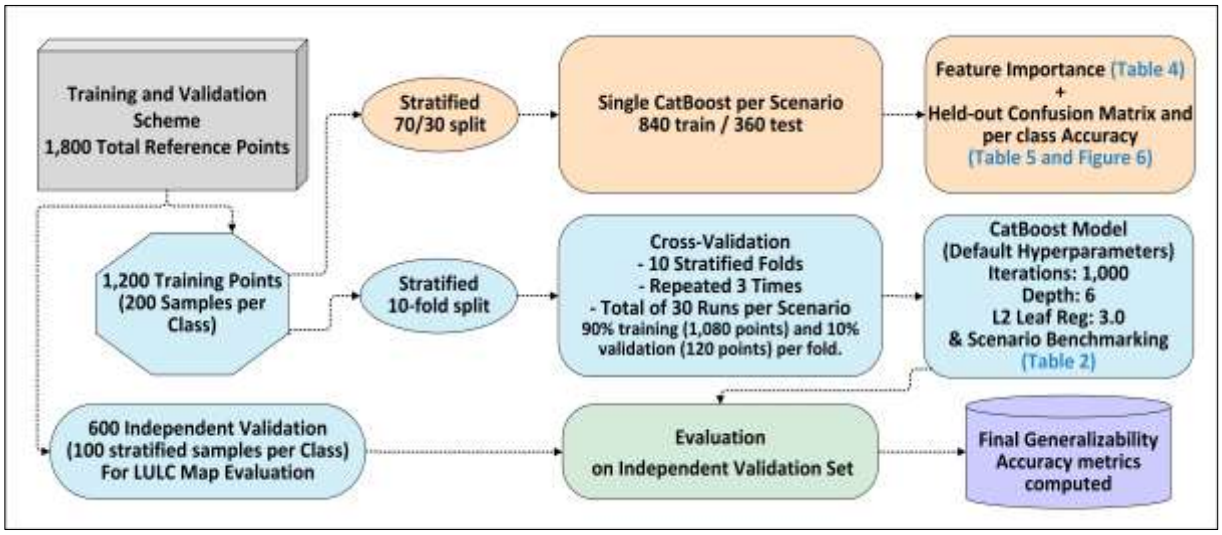


Figure 3. Workflow for CatBoost model training and independent validation

2.6 Accuracy assessment and statistical validation

2.6.1 Performance metrics

The performance of the classification system was evaluated using standard multi-class evaluation metrics that are generated from confusion matrices [3, 10]. Overall accuracy (OA) is defined as the ratio of correctly classified validation samples to the total number of samples across all six-land use land cover classes; producer's accuracy (PA) and user's accuracy (UA) determine class-specific omission and commission errors, respectively; and the F1 score is calculated as the harmonic mean of PA and UA. The Kappa coefficient is also utilized to adjust the measure of overall agreement based on expected chance agreement values, as defined in literature [3, 10, 30].

2.6.2 Scenario performance and improvement comparison

To assess the incremental accuracy contributions arising from the transitional addition of new sensor domains (Scenarios A to G), we developed two comparative enhancement indices. Let's define S_base as the basic model scenario (for instance, Scenario A using visible wavelengths only), and S_new as the subsequent, augmented feature scenario. The Absolute Accuracy Gain is a measure of the linear increase in accuracy (in percentage) between two modeling scenarios, which is calculated as follows:

$$\text{Gain}\% = OA(S_{\text{new}})\% - OA(S_{\text{base}})\% \ldots (8)$$

where:

$OA(S)\%$ is the overall accuracy of the scenario$S$ expressed as a percentage. This calculation can similarly be applied to compare class-specific F1 scores and Kappa coefficient values:

$$\text{Gain}_{\text{F1}}\% = F1(S_{\text{new}})\% - F1(S_{\text{base}})\% \ldots (9)$$

2.7 Statistical dispersion and model stability

In order to assess the stability, consistency, and reliability of the CatBoost classifier during the repeated 10-fold cross-validation approach (M = 30 independent runs per scenario), two strong estimators of statistical variation have been calculated: the standard deviation (SD) and the coefficient of variation (CV = SD/mean × 100%). A low value of SD

suggests that the classifier remains highly stable in its classification results, irrespective of the particular way in which the spatial reference dataset has been partitioned. Additionally, the coefficient of variation provides this measure in a way that is independent of the scale of the mean.

2.7.1 Non-parametric statistical benchmarking

Friedman test

Initially, the Friedman test was used to assess the universal null hypothesis [11], which states that all the seven progressive feature scenarios demonstrate similar performance distributions. In this case, n stands for the number of blocks (n=30 repeat cross-validation runs), and k stands for the number of scenarios used in the comparison (k=7).

The test statistic (χ_F^2) is calculated as:

$$\chi_F^2 = \frac{12}{n \cdot k(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \ldots (10)$$

where:

χ_F^2 is the Friedman test statistic, which asymptotically follows a chi-squared (χ^2) distribution with $k-1$ degrees of freedom;
$n$are the total number of blocks or repeated cross-validation runs ($n = 30$);
$k$is the total number of treatment groups or feature scenarios ($k = 7$);

$R_j$is the sum of ranks assigned to the $j$nth feature scenario across all n runs, defined as

$$R_j = \sum_{i=1}^{n} r_{ij} \ldots (11)$$

where:

$r_{ij}$is the rank (from 1 to k) of scenario j within the i-th cross-validation run.

Wilcoxon test

When the Friedman test rejected the global null hypothesis, the Wilcoxon signed-rank test [44] was then used to conduct pairwise comparisons between cases (such as scenario G versus scenario F). This test checks if the addition of the sensor types leads to a significant change in accuracy.

For any pairwise comparison across the n runs, let $d_i = x_{1,i} - x_{2,i}$ represent the difference in overall accuracy between two compared scenarios for run i. After excluding all cases where $d_i = 0$, the remaining N_r non-zero absolute differences $|d_i|$ are ranked from smallest (rank 1) to largest (rank N_r). The Wilcoxon test statistic (W) is defined as:

$$W = \min(W^+, W^-) \ldots (12)$$

where:
$W^+$is the sum of ranks connected with positive changes, representing runs where the expanded scenario outperformed the baseline.

$$W^+ = \sum_{i:d_i>0} \text{rank}(|d_i|) \ldots (13)$$

$W^-$is the sum of ranks associated with negative differences:

$$W^- = \sum_{i:d_i<0} \text{rank}(|d_i|) \ldots (14)$$

Kendall's coefficient of concordance (W)

Although the Friedman test clarifies whether there is a significant difference between the cumulative feature scenarios, and the Wilcoxon signed rank test helps figure out which two scenarios differ, both tests do not quantify the degree of agreement between different rankings of scenarios. For this purpose, the W-index was calculated [17], which is a measure of effect size and characterizes the level of concordance between the independent rankings of scenarios.

Out of the functions of W, we can derive Kendall's W from the Friedman rank-sum statistic as follows:

$$W = \chi^2 F / [n(k-1)] \ldots (15)$$

where:

$\chi^2 F$ is the Friedman test statistic obtained above;
n is the total number of blocks, or repeated cross-validation runs (n = 30);
k is the total number of treatment groups, or feature scenarios (k = 7).

McNemar's test

Since the predicted results for Scenario A (Baseline) and Scenario G (Best) were generated based on the same database of 600 independent validation points, comparing their overall accuracy to test the existence of statistical significance would not be appropriate, as it would consider their predictions as independent samples. Therefore, McNemar's test was carried out in order to evaluate if the two models exhibit different error behavior beyond what can be regarded as random variations of equally good classifiers [24].

For each of the 600 validation points, we constructed a 2 × 2 contingency table based on the classifications made by each model. The number of points correctly classified by Scenario A but wrongly classified by Scenario G is denoted as b, and the number of points incorrectly classified by Scenario A but correctly classified by Scenario G is denoted as c. McNemar's test statistic is given as:

$\chi^2 = (|b - c| - 1)^2 / (b + c)$ … (16)

2.8 SHAP model interpretability

The solutions to the problem of the ambiguity of ensemble tree-based classifiers were realized by employing Shapley additive explanations (SHAP) as the game-based technique that allows for decision reasoning and visualization of results [19]. SHAP uses an additive model of feature attribution, which enables obtaining an accurate degree of contribution of each of the input factors on the pixel-level predictions:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \dots (17)$$

where:

$g(z')$is the local explanatory surrogate model.
$z' \in \{0,1\}^M$represents the coalition vector of active (1) or inactive (0) features within a simplified local background space.
$M$is the maximum coalition size, representing the total number of input variables in the scenario.
ϕ_0 is the base value of the model, equivalent to the expected value of all predictions when no features are present.
$\phi_i \in$ the Shapley value (feature attribution score) represents the clean, additive contribution of the feature$i$ to the final prediction.

For each input variable, the unique Shapley value (ϕ_i) is mathematically formulated by evaluating the marginal contribution of feature i across all possible feature subsets [27]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)] \dots (18)$$

where:

F represents the complete set of all input variables utilized in the scenario.
Srepresents a sub-coalition (subset) of features drawn from F that excludes feature i.
$|S|$and$|F|$ represent the cardinalities (number of active features) of the subset$S$ and set F, respectively.
$f_x(S)$ is the prediction of the model for a specific target instance x when computed using only the features present in subset S.
$f_x(S \cup \{i\}) - f_x(S)$ is the marginal prediction contribution directly attributable to the addition of feature i.

3. Results and Discussion

3.1 Progressive scenario performance

Table 2 displays the 10-fold, three-times-repeated cross-validation performance (30 runs per stack) of CatBoost against the seven cumulative-feature stack levels. The OA increases steadily across progressively expanded stack sizes from 83.39% (SD = 3.73%) at Scenario A (Baseline, visual bands only) to 95.50% (SD = 1.49%) at the full 23-feature Scenario G stack; in addition to the absolute increase, the CV and 95% confidence interval half-width fall consistently from 4.48% (CV = 3.73/83.39 × 100) for the initial model to 1.56% for the full model, indicating dramatic model stabilization from the addition of feature domains with the resulting decrease in

predictive variance, again as predicted by ensemble learning theory (Figure 4) The largest single accuracy gain occurs from A (Baseline) to Scenario B (+6.14%), the addition of NIR/SWIR bands (B8, B11, and B12) reflecting a typical increase for arid/complex mountainous terrain; the visible bands are unable to distinguish similar bare land and forest land, which directly correlates with the ability of NIR/SWIR domains to detect water, vegetation status, and minerals of bare lands [36]. The next gain is from B to C (+1.36%); the thermal/LST model differentiates surface characteristics of vegetated and non-vegetated surfaces through their surface temperature related to vegetation canopy [35]. Two incremental increases from spectral indices added C to D (+0.36%) and from combined optical SAR texture added to the previous stack D to E (+0.42%) reveal minimal/insignificant gains, suggesting substantial data redundancy among NDVI, NDWI, NDBI, and BSI with input spectral bands as well as texturally derived optical and radar data. However, two substantial significant increases occurred with the inclusion of SAR backscatter from E to F (+1.53%) and topography (DEM, slope, and aspect) from F to G (+2.31%), the latter being the biggest single contribution aside from increased NIR/SWIR. This final addition topographic variables, which contain no purely spectral information corresponds to a statistically significant improvement in classification performance (Gain OA% = +2.31 percentage points; Wilcoxon signed-rank test, Holm-adjusted $p < 0.001$; Table 3), resolving approximately 34% of the remaining classification error present at the previous cumulative stack (relative error reduction = $(95.50 - 93.19) / (100 - 93.19) \times 100$). This improvement is consistent with topography aiding the discrimination of vegetated and non-vegetated classes that are vertically stratified by elevation in mountainous terrain, such as forested slope areas versus surrounding exposed surfaces. The runtimes varied from 3.4 min (Scenario A) to 18.9 min for the full Scenario G stack (Figure 4). The overlapping confidence intervals between 95% in Scenarios C-E imply the difference in accuracy among them is not statistically significant. Even with the additional cost of including special indexes and merged texture in Scenarios D and E compared to Scenarios C and B, respectively. However, Scenarios F and G demonstrate a statistically significant increase in accuracy together with a justified proportional increase in the run [1].

Table 2. Cross-validated CatBoost performance across the seven cumulative feature scenarios (10-fold × 3 repeats, n = 30 runs per stack).

Scenario	Feature	Mean OA%	SD%	CV%	CI-95%	F1%	Kappa	Gain OA%
A (Baseline)	3	83.39	3.73	4.48	±1.39	83.26	0.801	-
B	6	89.53	2.95	3.29	±1.10	89.50	0.874	+6.14
C	7	90.89	3.00	3.30	±1.12	90.84	0.891	+1.36
D	11	91.25	2.72	2.98	±1.02	91.24	0.895	+0.36
E	17	91.67	2.46	2.68	±0.92	91.60	0.900	+0.42
F	20	93.19	2.32	2.49	±0.87	93.12	0.918	+1.53
G	23	95.50	1.49	1.56	±0.56	95.48	0.946	+2.31

OA: Overall Accuracy, SD: Standard Deviation, CV: Coefficient of Variation, CI: Confidence Interval.

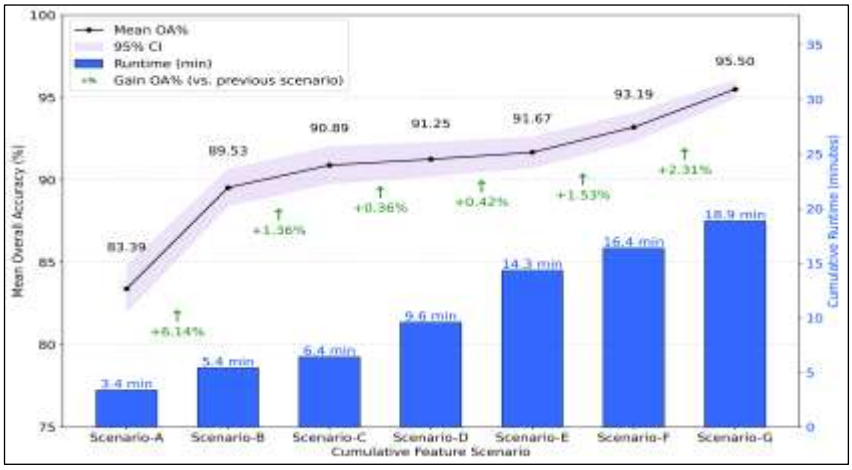


Figure 4. Mean overall accuracy (± 95% CI), cumulative runtime (minutes), and incremental gain across the seven cumulative feature scenarios.

3.2 Statistical significance of progressive fusion

A Friedman test on the results across all seven cumulative scenarios revealed a significant global difference in classification performance ($F = 131.11$, $p < 0.001$, $n = 30$ blocks, $k = 7$ treatments), rejecting the null hypothesis of no differences across scenarios' performance distributions.

The value of Kendall's coefficient of concordance ($W = 0.728$) suggests a good and consistent concordance among different blocks across scenarios regarding the relative ranking of classification performance, ruling out any suggestion that the improvement pattern we observed would be merely an artifact due to the luck of a small number of favorable folds. In addition, following pair-wise comparison (post hoc Wilcoxon signed-rank test; Table 3), we observe four significant sequential improvements among scenarios: Scenario A to B ($p = 0.000002$), B to C ($p = 0.001826$), E to F ($p = 0.000070$), and F to G ($p = 0.000010$). Non-significative pairwise transitions were observed from C to D (spectral indices, $p=0.105851$) and from D to E (textural indices, $p=0.373463$), resulting in agreement with the Small Gain OA% value from those pairs of scenarios presented on

Table 2, suggesting that those groups of features are not able to provide a distinct improvement over the immediately preceding stack composed, in this case, by the first stack including raw spectral and thermal information, although it is not intended to propose those groups of features as uninformative in an absolute sense. To control for family-wise Type I error across the six pairwise comparisons, Holm's step-down correction was applied to the Wilcoxon p-values (Table 3); all comparisons identified as significant under the uncorrected threshold ($\alpha = 0.05$) remained significant after Holm correction.

Table 3. Wilcoxon signed-rank test results for adjacent scenario comparisons.

Scenario comparison	W statistic	p-value	p-value-holm	Significant at $\alpha = 0.05$
Scenario B vs Scenario A	0.0	0.000002	0.000015	True
Scenario C vs Scenario B	59.5	0.001826	0.005477	True
Scenario D vs Scenario C	102.5	0.105851	0.211703	False
Scenario E vs Scenario D	140.5	0.373463	0.373463	False
Scenario F vs Scenario E	15.0	0.000070	0.000282	True
Scenario G vs Scenario F	9.0	0.000010	0.000048	True

3.3 Feature importance and SHAP interpretability for Scenario G

The performance of the 23-features full-scenario G model was estimated using CatBoost native importance, permutation importance, and mean absolute SHAP values with comparable rankings of the features (Table 4).

Topographically, the most informative features were DEM and SLOPE with native importance of 16.21% and 11.24% and contribution of 0.529 and 0.525, respectively. The largest increment in model accuracy from Baseline (A) to Scenario B (+6.14%) resulted from the addition of NIR/SWIR bands (B8, B11, and B12). The presented result highlights the high discriminative importance of the topographic features in the model because DEM and SLOPE contribute significantly to SHAP importance together.

These observations support the idea of elevation and slope being robust indicators of microclimatic and vegetation-zonation gradient in the mountains, which may increase the discrimination power of the model by separating different ecosystems.

Nevertheless, it needs to be mentioned that SHAP-based feature importance does not indicate any causal relation between the topography and climate variation, ecosystem distribution, or habitat destruction. B12 (SWIR2), NDVI, B4 (Red), and NDWI were the other most important features. SAR_VH and SAR_CONTRAST turned out to be informative for native and SHAP importance, thus highlighting the role of spectral and textural SAR features besides topographic and optical features. The strongest agreement between the measures was achieved by the native feature importance and SHAP for CatBoost (Spearman's $r = 0.96$; Kendall's $\tau = 0.858$). The permutation feature importance correlated less strongly with both native feature importance ($r = 0.79$) and SHAP ($r = 0.77$; $p < 0.001$ for all comparisons) [36].

However, the three metrics yielded a similar ordering for the importance of features, although not necessarily the same explanation regarding the contribution to the decrease in impurity, decrease in prediction accuracy, and Shapley value.

Topography was revealed to be the most important variable in explaining the land cover patterns, while vegetation and soil moisture indices were the most informative corroborating factors. Considering the high discriminatory power of textural SAR features for the regions of high topography and vegetation, especially Zagros Mountains, SAR should be included into the future multi-source models [37]. The overall highest increase was at the feature group level (Figure 5) by spectral bands (+6.14% OA), whereas elevation and SAR backscatter followed (+2.31% and +1.53%, respectively).

Thermal LST shows a small but significant increment (+1.36% OA), although the increment from spectral indices and aggregate GLCM/SAR texture is almost zero, +0.36% OA and +0.42% OA, respectively, despite six different textures that compose it [5, 38, 43].

The SHAP dependence analysis on the top six highest-ranked predictors produced a threshold type of behavior that is easily identifiable visually from the dependence plot (see Section 2.8) but not statistically through change point detection and was nevertheless physically interpretable as follows: the DEM value beyond which SHAP contribution to agricultural class prediction increases is around 800 meters in elevation (in line with the terracing practice located at lower elevations); SLOPE shows the opposite type of relationship where the SHAP contribution to open forest class prediction increases between 10° to 30° before leveling off; the SWIR2 band (B12) shows a separation point for built-up class above about 2,000 reflectance units; and NDVI shows separation of the water class from all other classes. In all cases, however, for the top 6 predictors, the relationship found through the SHAP dependence analysis is not operating alone as there exists an interacting physical variable that also represents each domain within the input space to the model and thus reinforces the evidence behind the use of multi-source

cumulative stacked features over any individual feature space from any one source.

The global predictor importance averaged over all six classes is shown in Table 4 below, which makes clear that the contribution of the feature to the prediction of any of the classes does not correspond to the ranking of the importance of the feature for any single one class in particular: the prediction for water classification is controlled primarily by the spectral and index features (NDVI, B8, B12, and NDWI) since water will have the spectral signature (low absorption) across all spectral bands and thus be highly sensitive to total brightness [22], that will also show up in the calculated index that will take very high values; the topographical features played hardly any role in this classification [25]. However, in the case of built-up class prediction, the most important variables are the topographic features DEM and a topographic feature SLOPE.

Table 4. Top 10 features under three importance-ranking methods (Scenario G, full 23 features).

Feature	Native		Permutation		SHAP	
	Rank	Importance	Rank	Importance	Rank	Importance
SLOPE	1	16.2106	1	0.0731	2	0.5254
DEM	2	11.2388	2	0.0458	1	0.5286
B12	3	8.9320	5	0.0139	3	0.3304
NDVI	4	8.0140	6	0.0117	4	0.3192
B4	5	6.9643	4	0.0147	5	0.2716
NDWI	6	6.1708	8	0.0089	6	0.2690
B11	7	5.2482	13	0.0003	7	0.2609
B8	8	5.0304	11	0.0031	10	0.2043
SAR_VH	9	4.6032	3	0.0267	8	0.2561
BSI	10	3.5326	12	0.0028	17	0.1262

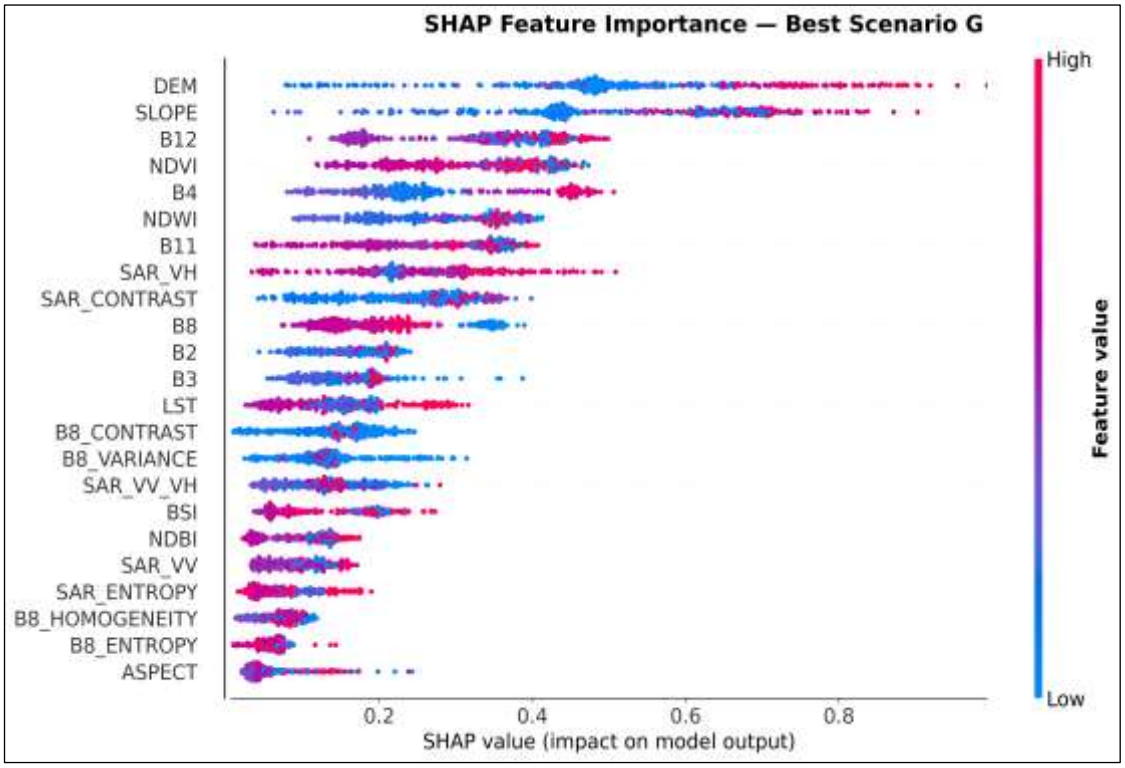


Figure 5. SHAP beeswarm summary plot for the Scenario G model. Each point shows a single validation sample's SHAP value per feature; color indicates feature value (red = high, blue = low), and horizontal position shows magnitude and direction of impact on the predicted class.

3.4 Per-class performance and confusion patterns (held-out test split)

Table 5 and Figure 6 show the best-performing Scenario G model; producer's and user's accuracies on the 30% test split were highest for spectrally and structurally distinct classes: water and built-up both achieved 100% PA/UA/F1, and agriculture achieved 100% PA with 98.4% UA. Confusion was greatest amongst the vegetated and soil land-cover classes at the drier extreme of the land-cover gradient. Dense forest (PA=91.7%, UA=96.5%) was infrequently misclassified as open forest (5/60 CV samples). Open forest (PA=81.7%, UA=83.1%) was subject to the largest number of confusions of any class, being misclassified as bareland in 9/60 sample points and dense forest in 2/60 sample points. Bareland (PA=90.0%, UA=85.7%) was confused reciprocally with open forest (5/60 samples) and on occasion with agriculture (1/60). The confusability of bareland and open forest is the predominant remaining error term for the model. While confusing, this is ecologically interpretable: this dryland oak forest readily degrades and transitions into soil/rock at canopy cover thresholds, a phenomenon readily observed and often documented in land degradation assessment [1, 33].

Table 5. Per-class accuracy for Scenario G model (30% test split)

Class	PA % (Recall)	UA % (Precision)	F1 %
Water	100.0	100.0	100.0
Built-up	100.0	100.0	100.0
Agriculture	100.0	98.4	99.2
Dense Forest	91.7	96.5	94.0
Open Forest	81.7	83.1	82.4
Bareland	90.0	85.7	87.8

Overall	93.9	93.9	93.9

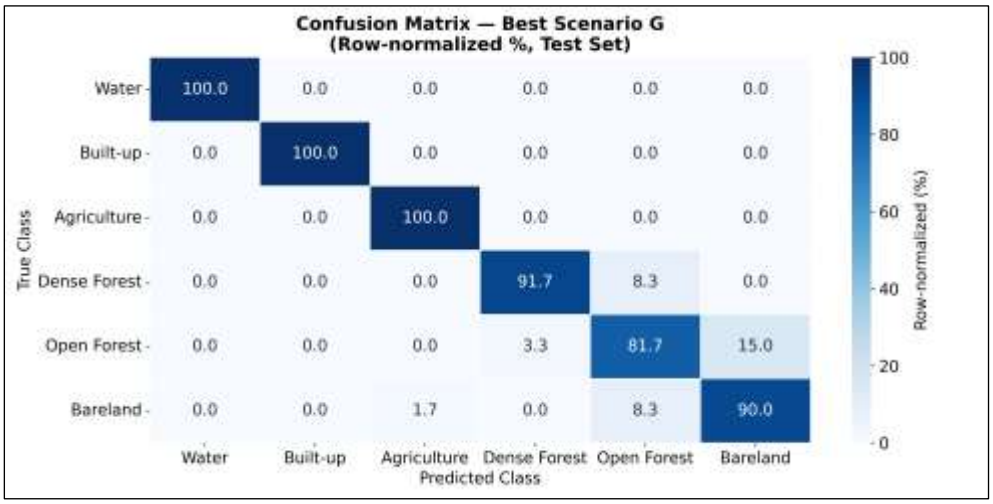


Figure 6. Confusion matrix for the Scenario G model, row-normalized percentages on the 30% test split.

3.5 Independent validation

Using the Scenario G approach to validate the classifier on 600 spatially independent validation points (100 per class), which have never been used for training, testing, or cross-validation (Figure 7). As a result, we get an accuracy of 95.33%, almost equivalent to the cross-validated average of 95.50% (the difference is just 0.17 percentage points) along with Kappa = 0.944. The high correlation of cross-validation results and spatially independent validation indicates that the reported accuracy was not artificially increased by the information flow between training folds. In addition, to decrease the spatial autocorrelation influence, the distance between training and validation points was fixed at least 300 m (see Section 2.3), which is significantly greater than the feature pixel size of 10 m and GLCM texture analysis window of 70 m, thus decreasing the chances of overlapping of spatial context for the training and validation sets. Thus, our approach meets the good practice guidelines of validating the classifier performance using a probability sample [16, 30].

From the data presented in Table 6, it can see that independent accuracy for each individual class followed a similar trend as the cross-validation and that it was in agreement with the CV confusion analysis: open forest (PA = 93.0%, UA = 91.2%) and bareland (PA = 91.0%, UA = 90.1%) were still the two classes having the worst performance, and the independent confusion matrix proved the very same axis of confusion for open forest-bareland confusion (7 points of open forest confused as bareland and 5 points of bareland confused as open forest). From this, it is evident that the confusion of open forest and bareland is a consistent pattern in both cross-validation and independent validation and, therefore, is not an artifact of CV sampling.

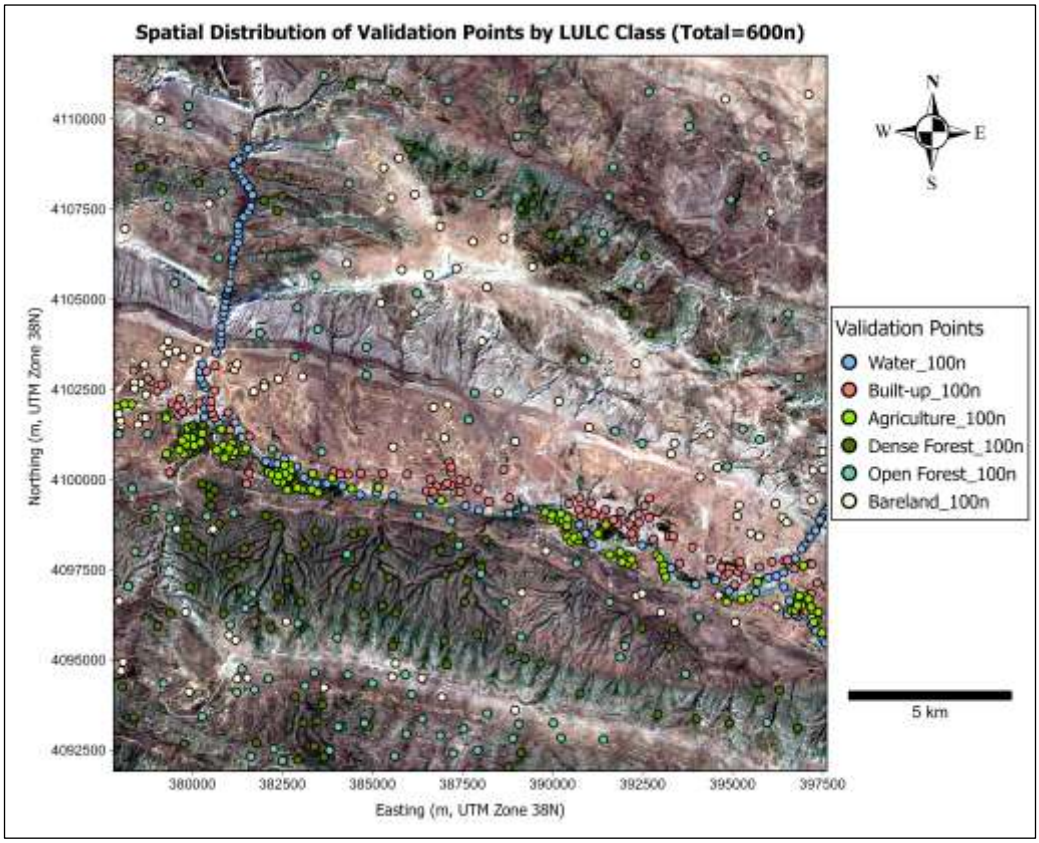


Figure 7. Spatial distribution of the 600 independent validation points across LULC classes

Table 6. Confusion Matrix of Independent validation accuracy (600 held-out points, Scenario G model).

Class	Water	Built-up	Agriculture	Dense Forest	Open Forest	Bareland	UA % (Precision)	F1 %
Water	99	0	1	0	0	0	100.0	99.5
Built-up	0	96	2	1	0	1	97.0	96.5
Agriculture	0	0	96	1	1	2	96.0	96.0
Dense Forest	0	0	0	97	3	0	98.0	97.5
Open Forest	0	0	0	0	93	7	91.2	92.1
Bareland	0	3	1	0	5	91	90.1	90.5
Overall	PA (Recall) = 95.33%						95.37	95.35

3.6 Spatial comparison mapping (Scenario A vs Scenario G)

Figures 8 and 9 display the final LULC maps generated using the base Scenario A and comprehensive Scenario G stack of 23 layers, respectively, for the 400 km² study area. Table 7 quantifies these results, revealing the landscape to be largely comprised of open forests (186.20 km²; 46.55% of the total mapped area), followed by barelands (119.40 km²; 29.85%), dense forests (67.40 km²; 16.85%), built-up areas (12.70 km²; 3.18%), agricultural lands (11.80 km²; 2.95 %), and water bodies (2.50 km²; 0.63%). These results appear to be compatible with the mountainous semi-arid landscape of the Amedi-Mergasur-Akre transect that is characterized by predominantly open and structurally degraded [1, 33]. The most obvious numerical disagreement is in the category of agriculture, which appears in more than the eleventh percent of the study area in the visually classified layer but less than the third in the one using the entire stack of images (Table 7). On the other hand, in the full-stack situation, the presence of bare lands and thick forests is common alongside relatively lower

land for agriculture and built-up areas. In all likelihood, these numbers indicate that the original visually classified image under-represented areas of bareland, which were identified through the use of the 23-band stack, which was able to learn to distinguish between areas of bareland and agriculture from their spectral differences in ways that the smaller set of imagery cannot. Since it can be shown that the accuracy of the Scenario G class reduces the remaining classification error by about 69.5% relative to the visually classified scheme ((95.33−84.67)/(100−84.67)×100), and since the point-based validation accuracy suggests that the classification of Scenario G is better aligned with ground observations at the validation points, although the validation is conducted through the use of 600 independent validation points rather than the entire study area this overprediction of agriculture could indeed be considered a policy-relevant result of using such narrow range imagery.

In order to prove that this discrepancy is due to accuracy rather than due to classification shift, both cases were evaluated against the same set of 600 independent validation points and then compared as done previously. Based on this held out dataset, for Scenario A, OA = 84.67% (Kappa = 0.816, Weighted F1 = 84.68%) while for Scenario G, OA = 95.33% (Kappa = 0.944, Weighted F1 = 95.35%), which shows a 10.66 percent-point improvement.

A McNemar test was performed, and these results are extremely statistically significant ($\chi^2 = 47.25$, $p < 0.001$), with Scenario G performing significantly better than the baseline. Table 8 shows the per-class accuracy analysis for the two scenarios; a difference is observed and localized to the affected classes, agriculture (PA increases from 76.0 to 96.0%) and bareland (PA increases from 77.0 to 91.0%), respectively. In comparing the mapping results with those of Scenario A (Baseline), there appears to be an apparent difference in the mapped surface, despite limiting the results to the visible band only. This is because of the total disagreement in terms of the pixels at 65.1%, meaning that 34.9% of the pixels are reclassified from one product to another (Figure 10). Agreement between the two maps was measured by calculating the percentage of the 10 m cells in the 400 km² study area for which both maps had the same land cover class at native resolution and the entire study area.

Table 7. Land cover area comparison between baseline (Scenario-A) and best (Scenario-G) classification scenarios

Class	Scenario-A (Baseline)		Scenario-G (Best)		Difference	
	Area (km²)	(%)	Area (km²)	(%)	Area (km²)	(%)
Water	8.00	2.00	2.50	0.63	-5.50	-1.38
Built-up	24.10	6.03	12.70	3.18	-11.40	-2.85
Agriculture	47.00	11.75	11.80	2.95	-35.20	-8.80
Dense Forest	49.60	12.40	67.40	16.85	+17.80	+4.45
Open Forest	178.60	44.65	186.20	46.55	+7.60	+1.90
Bareland	92.70	23.18	119.40	29.85	+26.70	+6.67
Total	400.00	100.00	400.00	100.00	0.00	0.00

Table 8. Per-class producer's (PA), user's (UA), and F1 accuracy for Scenario A and Scenario G evaluated on 600 independent validation points.

Class	Scenario A (Baseline)			Scenario G (Best)		
	PA%	UA%	F1%	PA%	UA%	F1%
Water	99.0	100.0	99.5	99.0	100.0	99.5
Built-up	85.0	86.7	85.9	96.0	97.0	96.5
Agriculture	76.0	79.2	77.6	96.0	96.0	96.0
Dense Forest	87.0	81.3	84.1	97.0	98.0	97.5
Open Forest	84.0	80.0	82.0	93.0	91.2	92.1
Bareland	77.0	81.1	79.0	91.0	90.1	90.5
Overall	84.67	84.72	84.68	95.33	95.38	95.35
Kappa	0.816			0.944		

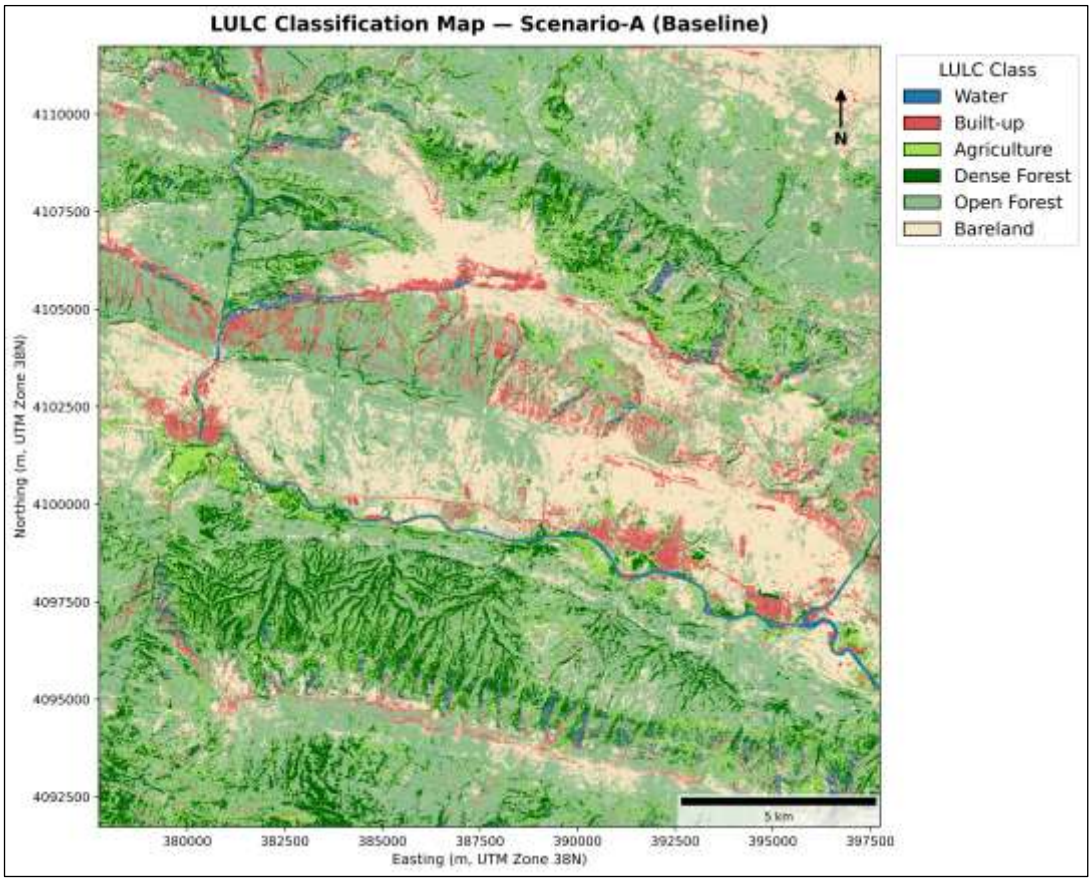


Figure 8. Classified LULC map for Scenario A (Baseline)

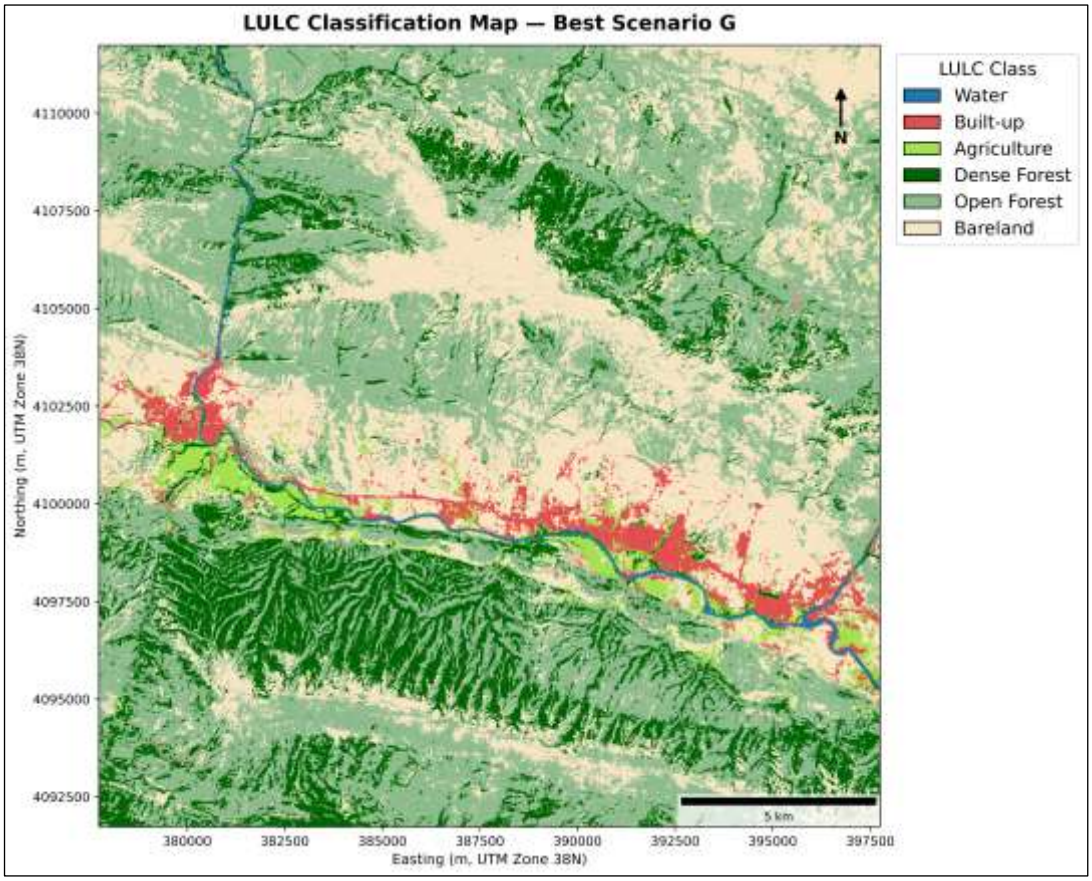


Figure 9. Classified LULC map for Scenario G (Best)

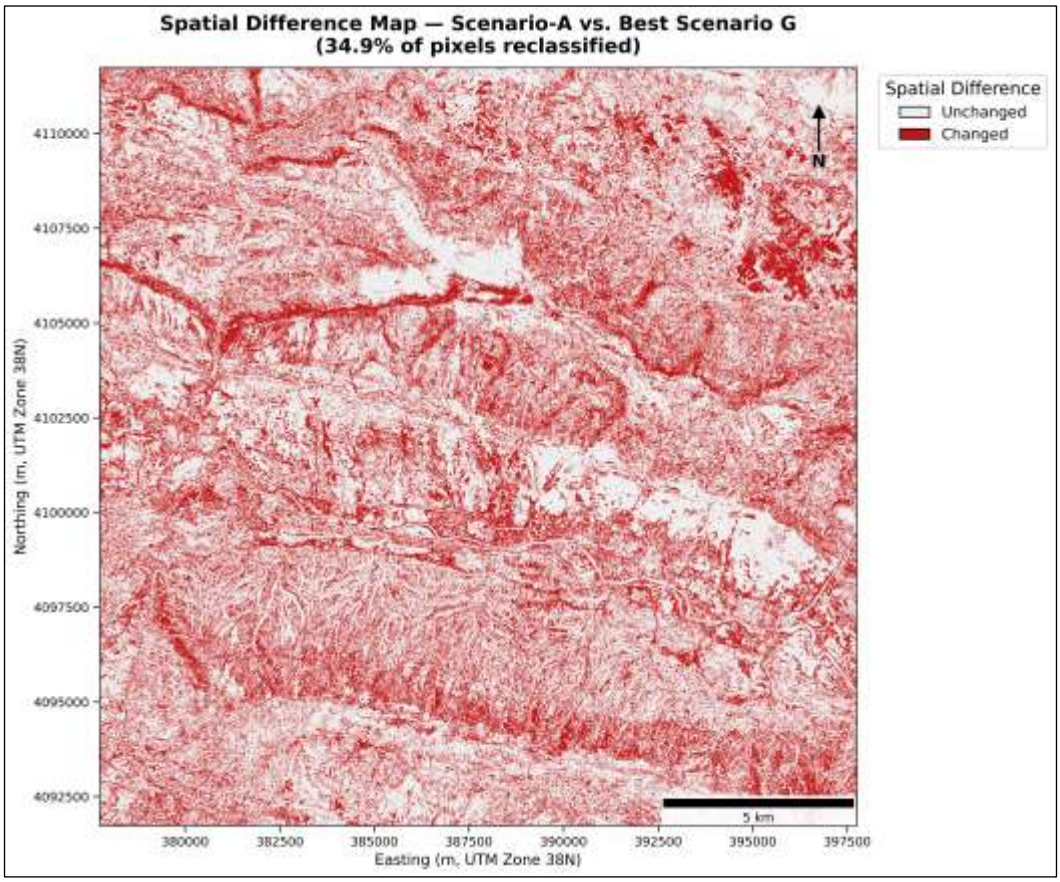


Figure 10. Spatial Difference Map for Scenarios A & G.

4. Conclusion

Within the scope of the datasets feature scenarios and the 400 km² study area evaluated here combining thermal, radar and topographic data led to a significant improvement in CatBoost-based land use and land cover classification. This was tested in the Amedi-Mergasur-Akre landscape of the Kurdistan Region, Iraq. When using visual bands (Scenario A) the cross-validated overall accuracy was 83.39%. When all 23 features were combined (Scenario G) accuracy improved to 95.50%.

The results were confirmed with a validation set of 600 points, where accuracy reached 95.33%. The biggest gains in accuracy came from infrared and shortwave infrared bands as well as from topographic variables. Because the feature scenarios were built in a way these improvements should be seen as marginal contributions added in sequence not as isolated effects from each individual data source. Slope and elevation consistently ranked among the important predictors across different importance measures native, permutation, and SHAP though high feature importance does not prove that these factors caused the accuracy gains.

Errors in classification often occurred between open-forest and bareland due to their spectral and structural similarities. While the study identified that a low-cost digital elevation model enhances accuracy, it did not test the confusion reasons. SAR data revealed insights missed by optical bands. Separating open forest from bareland remains challenging, suggesting future studies should use multi-temporal imagery and finer-scale textural features for better differentiation.

Data Availability Statement

All the used data in this study are available freely in the Google Earth Engine platform. Other elements such as the Python scripts and the Colab notebooks used to process and extract the data are accessible from the corresponding author upon reasonable demand.

Funding

This research was self-funded by the authors.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

R.K.M. was responsible for conceptualization, methodology, software, data collection, analysis, and validation. A.A.A was responsible for manuscript preparation, writing, review and editing. Authors have read and agreed to the submitted version of the manuscript.

References

[1] Alsinayi, N. B. M., M. H. O. Alzuhary, and H. A. A. Gaznayee. 2025. A comprehensive remote sensing approach for assessing forest degradation and environmental stress in Amedi District, Kurdistan Region of Iraq. Zanco Journal of Pure and Applied Sciences, 37(4): p. 46–69. https://doi.org/10.21271/ZJPAS.37.4.5.

[2] Benhssaine, K., J. Aabdousse, N. El Bouchti, H. Nechchadi, A. Ettaqy, Y. Abbas, and H. Berrougui. 2024. Anthropogenic activities and climate variability impacts on Juniperus thurifera forest of El Houanet, Morocco. Ecological Engineering & Environmental Technology, 25. http://dx.doi.org/10.12912/27197050/178278.

[3] Congalton, R. G.and K. Green. 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. 2 ed. Boca Raton: CRC Press. https://doi.org/10.1201/9781420055139.

[4] **Dagne, S. S., H. H. Hirpha, A. T. Tekoye, Y. B. Dessie, and A. A. Endeshaw. 2023.** Fusion of sentinel-1 SAR and sentinel-2 MSI data for accurate Urban land use-land cover classification in Gondar City, Ethiopia. Environmental Systems Research, 12(1): p. 40. https://doi.org/10.1186/s40068-023-00324-5.

[5] **De Luca, G., J. MN Silva, S. Di Fazio, and G. Modica. 2022.** Integrated use of Sentinel-1 and Sentinel-2 data and open-source machine learning algorithms for land cover mapping in a Mediterranean region. European Journal of Remote Sensing, 55(1): p. 52–70. https://doi.org/10.1080/22797254.2021.2018667.

[6] **Deressu, T. F., A. K. Bojer, and T. G. Debelee. 2025.** Enhancing land use and land cover classification with deep learning-based satellite imagery segmentation. International Journal of Applied Earth Observation and Geoinformation, 2025: p. 104839. https://doi.org/10.1016/j.jag.2025.104839.

[7] **Dobrinić, D., M. Gašparović, and D. Medak. 2021.** Sentinel-1 and 2 time-series for vegetation mapping using random forest classification: A case study of Northern Croatia. Remote Sensing, 13(12): p. 2321. https://doi.org/10.3390/rs13122321.

[8] **Farr, T. G., P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank, and D. Alsdorf. 2007.** The Shuttle Radar Topography Mission. Reviews of Geophysics, 45(2): p. RG2004–RG2004. https://doi.org/10.1029/2005RG000183.

[9] **Fatima, S., A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam. 2023.** Xgboost and random forest algorithms: an in depth analysis. Pakistan Journal of Scientific Research (PJOSR), 3(1): p. 26–31. https://doi.org/10.57041/vol3iss1pp26-31.

[10] **Foody, G. M. 2002.** Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80(1): p. 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4.

[11] **Friedman, M. 1937.** The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200): p. 675–701. https://doi.org/10.1080/01621459.1937.10503522.

[12] **Gaznayee, H. A. A., H. A. S. Razvanchy, F. Ahmad, and K. K. Hakzi. 2023.** Land Use Land Cover (LULC) Classification Using Google Earth Engine for the Kurdistan Region of Iraq Based on Sentinel-2: CultiVision for Research and Development. Erbil, Kurdistan Region of Iraq. https://iraq.un.org/en/271108-land-use-land-coverlulc-classification-kurdistan-region-iraq.

[13] **Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017.** Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202: p. 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

[14] **Horn, B. K. P. 1981.** Hill shading and the reflectance map. Proceedings of the IEEE, 69(1): p. 14–47. https://doi.org/10.1109/PROC.1981.11918.

[15] **Jiménez, F., G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles. 2017.** Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing, 234: p. 75–92.

https://doi.org/10.1016/j.neucom.2016.12.045.

[16] **Karasiak, N., J.-F. Dejoux, C. Monteil, and D. Sheeren. 2022.** Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. Machine Learning, 111(7): p. 2715–2740. https://doi.org/10.1007/s10994-021-05972-1.

[17] **Kendall, M. G.and B. Babington Smith. 1939.** The problem of m rankings. The Annals of Mathematical Statistics, 10(3): p. 275–287. https://doi.org/10.1214/aoms/1177732186.

[18] **Khwarahm, N. R., P. M. Najmaddin, K. Ararat, and S. Qader. 2021.** Past and future prediction of land cover and land-use change based on earth observation data by the CA–Markov model: a case study from Duhok Governorate, Iraq. Arabian Journal of Geosciences, 14(15): p. 1544–1544. https://doi.org/10.1007/s12517-021-07984-6.

[19] **Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020.** From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1): p. 56–67. https://doi.org/10.1038/s42256-019-0138-9.

[20] **Lundberg, S. M.and S.-I. Lee. Year.** A unified approach to interpreting model predictions. in Advances in Neural Information Processing Systems of Conference. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[21] **Malakar, N. K., G. C. Hulley, S. J. Hook, K. Laraby, M. Cook, and J. R. Schott. 2018.** An operational land surface temperature product for Landsat thermal data: Methodology and validation. IEEE Transactions on Geoscience and Remote Sensing, 56(10): p. 5717–5735. https://doi.org/10.1109/TGRS.2018.2824828.

[22] **Manyothwane, T.and G. Mengistu Tsidu. 2025.** Challenging global generalizations: superior land cover mapping in Botswana with a locally trained transformer model. Frontiers in Remote Sensing, 6: p. 1654692. https://doi.org/10.3389/frsen.2025.1654692.

[23] **McFeeters, S. K. 1996.** The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing, 17(7): p. 1425–1432. https://doi.org/10.1080/01431169608948714.

[24] **McNemar, Q. 1947.** Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2): p. 153–157. https://doi.org/10.1007/BF02295996.

[25] **Mertikas, S. P., P. Partsinevelos, C. Mavrocordatos, and N. A. Maximenko, 2021.** Environmental applications of remote sensing, in Pollution assessment for sustainable practices in applied sciences and engineering. Elsevier. p. 107–163. https://doi.org/10.1016/B978-0-12-809582-9.00003-7.

[26] **Nasir, S. M., K. V. Kamran, T. Blaschke, and S. Karimzadeh. 2022.** Change of land use/land cover in Kurdistan Region of Iraq:a semi-automated object-based approach. Remote Sensing Applications:Society and Environment, 26: p. 100713–100713. https://doi.org/10.1016/j.rsase.2022.100713.

[27] **Nohara, Y., K. Matsumoto, H. Soejima, and N. Nakashima. 2022.**

Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer methods and programs in biomedicine, 214: p. 106584. https://doi.org/10.1016/j.cmpb.2021.106584.

[28] Ogor, J. 2018. Design of algorithms for the automatic characterization of marine dune morphology and dynamics, ENSTA Bretagne-École nationale supérieure de techniques avancées Bretagne.https://doi.org/10.70675/ca675a65z85f3z4ac4zbf04za4d7a01f8f33.

[29] Okolie, C., J. Mills, A. Adeleke, J. Smit, and I. Maduako. 2023. The explainability of gradient-boosted decision trees for Digital Elevation Model (DEM) error prediction. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48: p. 161–168. https://doi.org/10.5194/isprs-archives-XLVIII-M-3-2023-161-2023.

[30] Olofsson, P., G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder. 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148: p. 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

[31] Onačillová, K., M. Gallay, D. Paluba, A. Péliová, O. Tokarčík, and D. Laubertová. 2022. Combining Landsat 8 and Sentinel-2 data in Google Earth Engine to derive higher resolution land surface temperature maps in urban environment. Remote Sensing, 14(16): p. 4076. https://doi.org/10.3390/rs14164076.

[32] Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Year. CatBoost: unbiased boosting with categorical features. in Advances in Neural Information Processing Systems of Conference. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c2486328554-Abstract.html.

[33] Rash, A., Y. Mustafa, and R. Hamad. 2023. Quantitative assessment of land use/land cover changes in a developing region using machine learning algorithms: a case study in the Kurdistan Region, Iraq. Heliyon, 9(11): p. e21253–e21253. https://doi.org/10.1016/j.heliyon.2023.e21253.

[34] Rasul, A., H. Balzter, G. R. Faqe Ibrahim, H. M. Hameed, J. Wheeler, B. Adamu, S. Ibrahim, and P. M. Najmaddin. 2018. Applying built-up and bare-soil indices from Landsat 8 to cities in dry climates. Land, 7(3): p. 81–81. https://doi.org/10.3390/land7030081.

[35] Rasul, A., S. Ibrahim, A. R. Onojeghuo, and H. Balzter. 2020. A trend analysis of leaf area index and land surface temperature and their relationship from a global to local scale. Land, 9(10): p. 388–388. https://doi.org/10.3390/land9100388.

[36] Rikimaru, A., P. S. Roy, and S. Miyatake. 2002. Tropical forest cover density mapping in Southeast Asia using World Forest Map. International Journal of Applied Earth Observation and Geoinformation, 4(1): p. 59–71. https://doi.org/10.1016/S0303-2434(02)00006-8.

[37] Schönbrodt-Stitt, S., N. Ahmadian, M. Kurtenbach, C. Conrad, N. Romano, H. R. Bogena, H. Vereecken, and P. Nasta. 2021. Statistical exploration of Sentinel-1 data, terrain parameters, and in-situ data for estimating the near-surface soil moisture in a mediterranean agroecosystem. Frontiers in Water, 3: p. 655837. https://doi.org/10.3389/frwa.2021.655837.

[38] Schulz, D., H. Yin, B. Tischbein, S. Verleysdonk, R. Adamou, and N.

Kumar. 2021. Land use mapping using Sentinel-1 and Sentinel-2 time series in a heterogeneous landscape in Niger, Sahel. ISPRS Journal of Photogrammetry and Remote Sensing, 178: p. 97–111. https://doi.org/10.1016/j.isprsjprs.2021.06.005.

[39] Sheykhmousa, M., M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni. 2020. Support vector machine versus random forest for remote sensing image classification: a meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: p. 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724.

[40] Torres, R., P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. N. Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, and F. Rostan. 2012. GMES Sentinel-1 mission. Remote Sensing of Environment, 120: p. 9–24. https://doi.org/10.1016/j.rse.2011.05.028.

[41] Tucker, C. J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8: p. 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.

[42] Voelsen, M., F. Rottensteiner, and C. Heipke. 2024. Transformer models for land cover classification with satellite image time series. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science. https://doi.org/10.1007/s41064-024-00299-7.

[43] Wang, X., Y. Gan, A. Iio, and Q. Wang. 2026. Enhancing seasonal leaf area index estimation using PLSR with Sentinel-2 spectral and texture features: the importance of feature engineering and feature selection. Precision Agriculture, 27(2): p. 42. https://doi.org/10.1007/s11119-026-10346-5.

[44] Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6): p. 80–83. https://doi.org/10.2307/3001968.

[45] Yang, Z., K. Zhou, and Y. Niu. 2026. Spatial distribution and environmental attributes dataset of China's large-scale data centers in 2024. Scientific Data. https://doi.org/10.1038/s41597-026-07548-x.

[46] Zha, Y., J. Gao, and S. Ni. 2003. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. International Journal of Remote Sensing, 24(3): p. 583–594. https://doi.org/10.1080/01431160304987.

[47] Zhao, S., K. Tu, S. Ye, H. Tang, Y. Hu, and C. Xie. 2023. Land use and land cover classification meets deep learning: a review. Sensors, 23(21): p. 8966–8966. https://doi.org/10.3390/s23218966.