

When Does Feature Engineering Matter? A Multi-Algorithm Comparison of Sentinel-2 Classifiers for Coastal and Mangrove Mapping in the Persian Gulf

Keivan Kabiri

kabiri@inio.ac.ir

Iranian National Institute for Oceanography and Atmospheric Science

Research Article

Keywords: Random Forest, Support Vector Machine, XGBoost, Maximum Likelihood Classification, Unmanned Aerial Vehicle (UAV) Validation, Land Use/Land Cover (LULC) Classification, Spectral Indices

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10973313/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Accurate land use/land cover (LULC) monitoring of mangrove and intertidal habitats along the Persian Gulf coast of Iran is essential for managing ecosystems that face both natural and industrial pressures, yet few studies have systematically compared classification algorithms for this region. This study compares four classifiers, Maximum Likelihood Classification (MLC), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), for classifying atmospherically corrected Sentinel-2 imagery into eight land cover classes across five coastal sites in southern Iran (Melgonzeh, Dayyer, Asalouyeh, Khour-e Azini, and Tiyaab). Each classifier was trained and tested (70/30 split) using two feature sets, six raw Sentinel-2 bands and the same bands augmented with four spectral indices, and results were cross-checked against high-resolution UAV orthomosaics acquired within 2–5 days of each Sentinel-2 scene. RF, SVM, and XGBoost achieved consistently high accuracy regardless of feature set (mean Overall Accuracy 94.9–96.9%), with only marginal, classifier-specific changes from adding indices (+ 0.9, – 1.0, and + 0.5 percentage points, respectively). MLC, in contrast, performed poorly with raw bands alone (mean Overall Accuracy 55.1%) but improved by an average of 27.3 percentage points (to 82.4%) once spectral indices were included, though it remained significantly less accurate than RF (McNemar's $\chi^2 = 165.885$, $p < 0.001$), whereas RF, SVM, and XGBoost did not differ significantly from one another. Built-up, Barren, and Soil classes were the most persistently confused group across all four classifiers. Independent UAV-based RGB congruence scores (44–49% on average) did not mirror this accuracy ranking, underscoring the value of combining ground-reference accuracy assessment with independent, training-free validation. These findings offer practical guidance for classifier selection in similar data-scarce, ecologically important coastal settings.

1. Introduction

Coastal zones occupy a small fraction of the Earth's surface yet sustain a disproportionate share of its biodiversity, fisheries, and human livelihoods, while remaining among its most dynamic and threatened environments (Green et al., 1996; Klemas, 2001). Among coastal habitats, intertidal mangrove ecosystems are particularly valuable: they stabilize shorelines, sequester carbon, and provide nursery habitat for numerous marine species, yet they are also highly sensitive to both natural stressors, such as salinity change and sea-level rise, and anthropogenic pressures, including land reclamation, aquaculture expansion, and industrial development. The inherent complexity of these transitional environments has long motivated efforts to develop systematic frameworks for characterizing them, whether through geomorphological classification schemes (Finkl, 2004) or, increasingly, through remote sensing-based land use/land cover (LULC) mapping (Kesgin & Nurlu, 2009; Shalaby & Tateishi, 2007).

Because coastal habitats are frequently extensive, tidally dynamic, and logistically difficult to survey on the ground, remote sensing has become the primary means of monitoring their extent and condition, with applications ranging from benthic habitat and bathymetry mapping in clear tropical waters (Louchard et al., 2003) to landscape-level indicators of estuarine and coastal habitat change (Klemas, 2011). Among available sensors, the European Space Agency's Sentinel-2 mission has emerged as a

particularly attractive data source for coastal LULC classification specifically, owing to its 10 m spatial resolution, five-day revisit interval, red-edge bands, and free and open data policy (Phiri et al., 2020). Sentinel-2 imagery has since supported land cover mapping across diverse coastal settings, from saltmarsh and sand dune habitats characterized using multi-source elevation data (Brown, 2004) to intertidal vegetation exhibiting pronounced seasonal phenology (Légaré et al., 2022), and it has increasingly been paired with feature-optimization and machine-learning approaches to improve accuracy in spectrally complex, high-resolution coastal scenes (Chen et al., 2021).

Despite this growing body of work, remote sensing-based LULC classification along the southern Iranian coastline remains comparatively limited, with existing Sentinel-2 studies concentrated heavily on the mangrove ecosystems of Qeshm Island (Ghorbanian et al., 2021; Soffianian et al., 2023; Bihamta Toosi et al., 2020), leaving other ecologically significant sites less thoroughly characterized. The present study addresses five such sites along the Persian Gulf coast of southern Iran—Melgonzeh, Dayyer, Asalouyeh, Khour-e Azini, and Tiyaab—each supporting mangrove or intertidal habitat, with several also subject to increasing pressure from coastal development and, in the case of Asalouyeh, large-scale oil and gas infrastructure. Site-scale UAV surveys conducted across this broader region (Kabiri, 2020; Miraki et al., 2023; Kabiri & Abedi, 2025) have demonstrated the value of very high-resolution drone imagery for mapping coastal features in fine detail, but such surveys cover only limited extents and cannot be repeated routinely at a regional scale, underscoring the continued need for reliable satellite-based classification methods, and for independent means of validating them.

A parallel line of research has compared classification algorithms for extracting LULC information from multispectral imagery. Maximum Likelihood Classification (MLC), a long-established parametric method, has been widely applied to coastal mapping problems, including land cover change detection along the Mediterranean coast of Egypt (Shalaby & Tateishi, 2007) and UAV-based coastal ecosystem mapping in the Persian Gulf (Kabiri, 2020). More recently, non-parametric machine-learning classifiers, particularly Support Vector Machine (SVM) and Random Forest (RF), have generally matched or outperformed MLC in Sentinel-2-based studies, including several targeting mangrove ecosystems in the region specifically (Ghorbanian et al., 2021; Miraki et al., 2023; Soffianian et al., 2023). Extreme Gradient Boosting (XGBoost), an ensemble method still less commonly evaluated in remote sensing applications, has nonetheless shown accuracy competitive with, or superior to, SVM, RF, and even deep-learning models in comparative studies from boreal (Abdi, 2020) and mangrove (Seydi et al., 2024) landscapes. Deep-learning approaches such as convolutional neural networks have also been applied to coastal landscape classification with high reported accuracy (Giang et al., 2023; Liu et al., 2024), though these typically demand substantially larger labeled training datasets than are practically available for most regional coastal studies. Notably, however, few studies have systematically compared MLC, SVM, RF, and XGBoost together within a single, consistent framework across multiple sites, evaluating both raw spectral bands and index-augmented feature sets, and none, to our knowledge, have done so along the Persian Gulf coast of Iran.

A further limitation concerns validation strategy. Training samples for supervised classification are commonly compiled from field surveys or high-resolution reference imagery such as Google Earth (Eskandari et al., 2020), but independent validation against very high-resolution reference data is far less common, and rarer still is such validation used to compare multiple classifiers against one another, rather than to confirm the performance of a single chosen method. UAV imagery, with its centimeter-scale resolution, offers exactly this kind of independent, temporally matched reference; UAV-based mapping has already proven valuable for characterizing coastal ecosystems at the site scale in this region (Kabiri, 2020; Miraki et al., 2023), yet it has rarely been paired with multi-classifier Sentinel-2 comparisons as an additional accuracy cross-check.

Building on this body of work, the present study aims to systematically compare four classification algorithms, MLC, SVM, RF, and XGBoost, for the LULC classification of Sentinel-2 imagery across five coastal sites in southern Iran: Melgonzeh, Dayyer, Asalouyeh, Khour-e Azini, and Tiyyab. For each site, a Sentinel-2 scene was selected to minimize cloud cover while also minimizing the temporal offset from a corresponding high-resolution UAV survey conducted between 2020 and 2023. Beyond comparing classifier accuracy across raw bands and spectral index combinations, this study applies McNemar's test to evaluate the statistical significance of pairwise accuracy differences and uses high-resolution UAV-derived orthomosaics as an independent reference against which satellite-based classification outputs are visually and quantitatively cross-checked. In doing so, this research seeks to identify the most robust and transferable classification approach for coastal and mangrove-associated LULC mapping in the Persian Gulf region, while offering one of the first systematic, multi-site, multi-algorithm evaluations of Sentinel-2 classification performance for this ecologically important yet understudied coastline.

2. Materials and Methods

2.1. Study Area

The study area comprises five coastal sites along the northern coast of the Persian Gulf, southern Iran, spanning approximately 51°25'–57°10'E and 26°20'–28°00'N across Bushehr and Hormozgan provinces (Fig. 1). The region has an arid, subtropical climate with hot summers and mild winters, and experiences a mixed, predominantly semi-diurnal tidal regime that exposes extensive intertidal mudflats at low tide (Afshar-Kaveh et al., 2020). Its coastline is characterized by these tidal flats, sabkha, and patches of mangrove forest dominated by *Avicennia marina*, interspersed with areas of growing coastal development.

Three of the sites, Melgonzeh, Dayyer, and Asalouyeh (Fig. 1a–c), lie in Bushehr Province, in the vicinity of Nayband Bay. Melgonzeh and Dayyer retain relatively undisturbed mangrove and intertidal habitat, whereas Asalouyeh, despite hosting its own small mangrove stand, is also home to the onshore gas-processing and petrochemical facilities of the South Pars Special Economic Energy Zone, making it a rare case of mangrove habitat under direct industrial pressure (Kabiri, 2020). The remaining two sites,

Khour-e Azini and Tiyab (Fig. 1d–e), lie roughly 400 km to the southeast in Hormozgan Province, near Sirik and Minab. Khour-e Azini is notable for hosting *Rhizophora mucronata*, a mangrove species otherwise rare along the Iranian coast (Miraki et al., 2023), while Tiyab supports an extensive mangrove stand at the eastern margin of the study area.

All five sites were selected because each had previously been surveyed by UAV between 2020 and 2023, yielding high-resolution orthomosaics that, although covering a smaller extent than the corresponding Sentinel-2 scenes, provide an independent reference for evaluating classification accuracy (Kabiri, 2020; Miraki et al., 2023). Together, the sites span a gradient from near-natural mangrove and intertidal ecosystems to areas under substantial industrial influence, offering an ecologically diverse basis for comparing classification algorithms across the Persian Gulf coast of Iran (Ghorbanian et al., 2021; Soffianian et al., 2023; Bihamta Toosi et al., 2020). Across all five sites, the classification scheme adopted in this study comprises eight land cover classes representative of these coastal and intertidal environments: mangrove, mudflat, soil, shallow water, deep water, barren, built-up, and vegetation.

2.2. Satellite and UAV Imagery, and Training/Validation Data

For each of the five study sites, a single Sentinel-2 Level-1C scene was selected based on minimal cloud cover and the shortest possible temporal offset from the corresponding UAV survey; the resulting offsets ranged from only 2 to 5 days across sites (Table 1). Because Level-1C products represent top-of-atmosphere reflectance rather than surface reflectance, all scenes were atmospherically corrected using the ACOLITE (version 20260421.0) processor (Vanhellemont & Ruddick, 2018). ACOLITE's dark spectrum fitting algorithm was developed specifically for aquatic and coastal environments and has been shown to outperform standard, land-oriented atmospheric correction approaches under nearshore and turbid-water conditions, making it well suited to the intertidal and coastal settings examined in this study (Warren et al., 2019; Kabiri, 2025).

At each site, a high-resolution UAV orthomosaic was available from a dedicated survey conducted as part of earlier fieldwork carried out between 2020 and 2023 (Table 1). Although each orthomosaic covers a smaller extent than the corresponding Sentinel-2 scene, its centimeter-scale resolution allows individual land cover features to be identified with confidence; a thumbnail of each orthomosaic is included in the last column of Table 1 for visual reference. Beyond illustrating the study sites, these orthomosaics later serve as an independent visual and quantitative check on the Sentinel-2-based classification results.

Ground reference data for training and validating the classification algorithms were compiled from two complementary sources. First, during the same fieldwork campaigns, ground observations were recorded at each site, including visual assessment of land cover and ground-level photographs with recorded GPS coordinates for specific features of interest, such as individual mangrove (*Avicennia marina*) trees. Second, additional training and test samples, covering land cover types not fully sampled in the field, were obtained through visual interpretation of high-resolution Google Earth imagery, a widely

used approach for generating reference data where systematic field sampling of every cover type is impractical (Eskandari et al., 2020). The field-based and Google Earth-derived samples were combined to form the training and validation datasets used consistently across all four classification algorithms (MLC, SVM, RF, and XGBoost) at each site.

2.3. Feature Extraction and Classification Algorithms

Following atmospheric correction with ACOLITE (Section 2.2), six reflectance bands were extracted from each Sentinel-2 scene: the blue (B2), green (B3), and red (B4) visible bands, the near-infrared band (B8), and the two shortwave-infrared bands (B11 and B12). Bands were identified programmatically by central wavelength to accommodate minor naming differences between ACOLITE outputs for Sentinel-2A and Sentinel-2B, and the two 20 m shortwave-infrared bands were resampled to match the 10 m grid of the visible/near-infrared bands.

In addition to these six raw bands, four spectral indices were calculated to enhance the separability of key coastal land cover classes (Table 2): the Normalized Difference Vegetation Index (NDVI; Rouse et al., 1974) to highlight vegetated and mangrove surfaces; the Modified Normalized Difference Water Index (MNDWI; Xu, 2006) and the shadow-optimized variant of the Automated Water Extraction Index (AWEI_sh; Feyisa et al., 2014) to enhance discrimination of open water from shadow and other dark surfaces; and the Normalized Difference Turbidity Index (NDTI; Lacaux et al., 2007) to capture the contrast between clearer and more turbid or sediment-laden water and mudflat surfaces. The six raw bands and four indices were stacked into a ten-band GeoTIFF for each scene, forming the two feature sets compared during classification.

Table 2

Spectral indices computed from Sentinel-2 reflectance bands (B2 = blue, B3 = green, B4 = red, B8 = near-infrared, B11 = shortwave-infrared 1, B12 = shortwave-infrared 2) and added to the six raw bands to form the Raw + Indices feature set.

Index	Formula	Purpose	Reference
NDVI	$NDVI = \frac{B_8 - B_4}{B_8 + B_4}$	Highlights vegetated and mangrove surfaces	Rouse et al. (1974)
MNDWI	$MNDWI = \frac{B_3 - B_{11}}{B_3 + B_{11}}$	Enhances discrimination of open water	Xu (2006)
$AWEI_{sh}$	$AWEI_{sh} = B_2 + 2.5B_3 - 1.5(B_8 + B_{11}) - 0.25B_{12}$	Enhances water discrimination from shadow and dark surfaces	Feyisa et al. (2014)
NDTI	$NDTI = \frac{B_4 - B_3}{B_4 + B_3}$	Captures turbidity contrast in water and mudflat surfaces	Lacaux et al. (2007)

Prior to classification, a valid-data mask was applied to exclude no-data and background pixels, those where all six raw reflectance bands were uniformly zero or undefined, from both training and prediction. Each remaining pixel was assigned to one of eight land cover classes representative of the study region's coastal and intertidal environments: Mangrove, Mudflat, Soil, Shallow Water, Deep Water, Barren, Built-up, and Vegetation.

Four classification algorithms, implemented in Python using scikit-learn (Pedregosa et al., 2011) and XGBoost (Chen & Guestrin, 2016), were compared at each site. Maximum Likelihood Classification (MLC) was implemented as a regularized Quadratic Discriminant Analysis (regularization parameter = 0.1); under the assumption of multivariate-normal, class-specific covariance structures, QDA is mathematically equivalent to the classical Gaussian maximum-likelihood classifier long used in remote sensing software (Gewali et al., 2018). Support Vector Machine (SVM) classification used a radial basis function kernel ($C = 10$, $\gamma = \text{"scale"}$; Cortes & Vapnik, 1995). Random Forest (RF) was implemented as an ensemble of 200 decision trees (Breiman, 2001). Extreme Gradient Boosting (XGBoost) was configured with 150 boosting rounds and a learning rate of 0.1 (Chen & Guestrin, 2016). A fixed random seed was used throughout to ensure reproducibility.

At each site, reference pixels drawn from the training/validation polygons (Section 2.2) were split into training (70%) and test (30%) subsets by stratified random sampling to preserve class proportions. To evaluate whether the addition of spectral indices improved classification performance, each of the four classifiers was trained and tested twice per site: once using only the six raw reflectance bands (Raw

Bands Only) and once using the full ten-band stack of raw bands plus derived indices (Raw + Indices), yielding eight classifier–feature-set combinations evaluated independently for each of the five study sites.

Classification performance was assessed using Overall Accuracy and the Kappa coefficient (Cohen, 1960; Congalton, 1991), computed from the held-out test set for every combination of site, feature set, and classifier. For each site, a final thematic land cover map was produced using the Raw + Indices feature set, predicting the class of every valid pixel across the full scene, and exported as a color-coded GeoTIFF for visual comparison and, subsequently, cross-checking against the corresponding UAV orthomosaic (Section 2.2).

To assess whether observed accuracy differences were statistically significant rather than incidental, pairwise McNemar's tests (McNemar, 1947) were conducted on the pooled test-set predictions across all five sites, following standard practice for classifier comparison in remote sensing (Foody, 2004). Random Forest was used as the reference classifier and compared against SVM, XGBoost, and MLC using their Raw + Indices predictions; an additional McNemar's test compared Random Forest's own pooled predictions between the Raw Bands Only and Raw + Indices feature sets to test whether the inclusion of spectral indices produced a statistically significant change in accuracy.

2.4. Independent Validation Using UAV-Derived RGB Imagery

As an additional, independent check on classification quality, each site's classified Sentinel-2 map (produced using the Raw + Indices feature set; Section 2.3) was compared against its corresponding UAV RGB orthomosaic (Section 2.2). Because the UAV imagery was not used in training or testing any classifier, this comparison provides a check on classification plausibility that is fully independent of the reference polygons used for accuracy assessment.

For each site, pixels where all three UAV RGB channels were simultaneously near-zero (≤ 5 , unphotographed black borders) or near-saturated (> 250 , white borders) were excluded as no-data. The UAV RGB imagery was then resampled from its native centimeter-scale resolution to the 10 m Sentinel-2 grid using area-weighted averaging, so that each Sentinel-2 pixel was assigned the mean RGB value of all UAV pixels falling within its footprint; the corresponding valid-data mask was resampled using nearest-neighbor interpolation to preserve its binary character. From the resampled RGB values, four descriptors were derived for every valid pixel: the Excess Green Index ($\text{ExG} = 2G - R - B$; Woebbecke et al., 1995), relative green and relative blue (each band's share of total $R + G + B$ brightness), and overall brightness (the mean of R, G, and B).

Each pixel's classified land cover label was then evaluated against these descriptors using a set of physically motivated rules, one per class or class group (Table 3). Because Soil, Barren, and Built-up share the same broad spectral signature in RGB imagery, they were evaluated against a single shared rule rather than three independent ones; the resulting congruence score for these classes therefore

reflects consistency with a general bright, non-water terrestrial surface rather than discrimination among the three classes individually.

Table 3
Physical/spectral rules applied to the resampled UAV RGB values to assess congruence with each classified land cover label.

Class(es)	Physical Rule (based on resampled UAV RGB)
Mangrove, Vegetation	ExG > 10, and Green > Red and Green > Blue
Mudflat	Brightness between 40 and 150, and Red – Green < 30
Soil, Barren, Built-up	Brightness > 90, and Red ≥ Blue
Shallow Water	Blue > Red, Relative Green > 0.30, and Brightness between 40 and 170
Deep Water	Blue > Green > Red, Relative Blue > 0.40, and Brightness < 80

For each classified map, the percentage of valid pixels satisfying the rule appropriate to their assigned class was calculated as an overall RGB congruence score, and these scores were compared across the four classification algorithms (MLC, SVM, RF, and XGBoost) at each site. This congruence score is intended as a complementary plausibility check based on independent, high-resolution imagery, rather than a substitute for the Overall Accuracy and Kappa coefficient reported in Section 2.3, which remain the primary accuracy metrics derived from the labeled reference data.

3. Results

3.1. Visual Comparison of Classified Maps

The classified land cover maps produced by all four algorithms across the five study sites, generated from the Raw + Indices feature set (Section 2.3), are shown in Fig. 2. For visual clarity only, a 3×3 majority filter was applied to reduce salt-and-pepper noise in these maps; this filter was not applied to the pixel values used for the accuracy assessment (Section 3.2) or the UAV congruence analysis (Section 3.3), so Fig. 2 appears somewhat smoother than the data underlying the numerical results reported below.

At most sites, the maps produced by Random Forest, SVM, and XGBoost are visually very similar, correctly delineating the general spatial pattern of mangrove, mudflat, water, and terrestrial classes. MLC maps, by contrast, show visibly greater spatial noise and characteristic large-area misclassifications at some sites: at Asalouyeh, MLC classifies large portions of the bay interior as deep rather than shallow water in a speckled pattern not present in the other three maps, and at Khour-e Azini, MLC assigns a much larger contiguous area to the built-up class than the other classifiers, at the expense of the barren and soil classes visible in the Random Forest, SVM, and XGBoost outputs. These qualitative differences are consistent with the substantially lower quantitative accuracy of MLC reported in Section 3.2.

3.2. Classification Accuracy Assessment

Overall Accuracy (OA) and the Kappa coefficient for each site, feature set, and classifier are summarized in Table 4 and Fig. 3. Random Forest, SVM, and XGBoost performed consistently well across all five sites and both feature sets, with mean OA of 95.5%, 96.9%, and 94.9%, respectively, using the raw Sentinel-2 bands alone, and 96.4%, 95.9%, and 95.4% when spectral indices were added (Table 4). MLC, in contrast, performed markedly worse with the raw bands alone (mean OA = 55.1%, ranging from 47.4% at Tiyaab to 62.8% at Melgonzeh) but improved substantially with the addition of spectral indices (mean OA = 82.4%, ranging from 75.7% at Tiyaab to 87.5% at Dayyer), an average gain of 27.3 percentage points that was consistent in direction and magnitude across all five sites (range: +23.8 to + 31.9 points; Fig. 3). By comparison, adding indices changed Random Forest's OA by a small, consistently positive + 0.9 percentage points on average, left XGBoost's OA essentially unchanged (+ 0.5 points on average, with small gains at some sites and small losses at others), and slightly reduced SVM's OA at four of the five sites (by up to 2.3 points), for an average change of – 1.0 percentage points. Kappa coefficients followed the same pattern as OA at every site and are reported alongside OA in Table 4.

Table 4
Overall Accuracy (%) and Kappa coefficient (in parentheses) for each classifier, study site, and feature set.

Site	Feature Set	Random Forest	SVM	XGBoost	MLC
Asalouyeh	Raw Bands Only	96.86 (0.964)	99.48 (0.994)	96.86 (0.964)	57.59 (0.500)
Asalouyeh	Raw + Indices	97.38 (0.970)	98.95 (0.988)	97.91 (0.976)	82.20 (0.793)
Dayyer	Raw Bands Only	97.92 (0.976)	97.92 (0.976)	95.14 (0.944)	59.72 (0.532)
Dayyer	Raw + Indices	98.61 (0.984)	97.92 (0.976)	94.44 (0.936)	87.50 (0.857)
Khour-e Azini	Raw Bands Only	93.52 (0.924)	93.06 (0.920)	93.52 (0.924)	48.15 (0.373)
Khour-e Azini	Raw + Indices	94.44 (0.935)	91.20 (0.898)	93.52 (0.925)	80.09 (0.764)
Melgonzeh	Raw Bands Only	93.02 (0.919)	97.09 (0.966)	93.60 (0.926)	62.79 (0.558)
Melgonzeh	Raw + Indices	93.60 (0.926)	94.77 (0.939)	95.35 (0.946)	86.63 (0.844)
Tiyab	Raw Bands Only	96.36 (0.958)	97.17 (0.967)	95.55 (0.948)	47.37 (0.372)
Tiyab	Raw + Indices	97.98 (0.977)	96.76 (0.963)	95.95 (0.953)	75.71 (0.715)

The confusion matrices for the Raw Bands Only and Raw + Indices feature sets (Figs. 4 and 5, respectively) show where classification errors are concentrated. With raw bands alone, Random Forest, SVM, and XGBoost separated all eight classes well, with per-class producer's accuracies mostly above 90%; the main source of error for all three was confusion between Built-up and Barren (7.7–9.2% of Built-up pixels misclassified as Barren). MLC, by contrast, showed widespread confusion across nearly every class pair with raw bands alone: only 29.4% of Mangrove pixels, 46.7% of Soil pixels, 49.1% of Shallow Water pixels, and 21.6% of Deep Water pixels were correctly labeled, with Deep Water frequently confused even with spectrally dissimilar classes such as Mangrove (20.5%) and Mudflat (14.8%). Adding spectral indices sharply improved MLC's per-class accuracy for exactly the classes the indices were designed to target: Mangrove rose to 94.1% correct, Shallow Water to 96.6%, and Deep Water to 71.6% (Fig. 5), consistent with the role of NDVI in separating vegetated classes and MNDWI/AWEI_sh in separating water-depth classes (Section 2.3). Across all four classifiers, however, Built-up, Barren, and Soil remained the most persistently confused group even with indices included, with Built-up correctly classified in only 55.6–86.6% of cases depending on the classifier, reflecting the limited spectral separability of these three bright, dry terrestrial classes at Sentinel-2's 10 m resolution.

Pairwise McNemar's tests on the Raw + Indices predictions, pooled across all five sites, confirmed that the accuracy difference between Random Forest and MLC was highly significant ($\chi^2 = 165.885$, $p < 0.001$; Table 5), consistent with the large gap in OA and Kappa reported above. In contrast, Random Forest did not differ significantly from SVM ($\chi^2 = 0.431$, $p = 0.512$) or from XGBoost ($\chi^2 = 3.349$, $p = 0.067$), indicating that the small differences in OA among these three classifiers are not statistically distinguishable at $\alpha = 0.05$. The comparison of Random Forest's own predictions between the Raw Bands Only and Raw + Indices feature sets was also statistically significant ($\chi^2 = 3.903$, $p = 0.048$), indicating that even Random Forest's modest, + 0.9 percentage point mean improvement from adding spectral indices reflects a systematic rather than incidental change in per-pixel predictions.

Table 5

Pairwise McNemar's test results comparing pooled test-set predictions (across all five sites) between classifiers on the Raw + Indices feature set, and between feature sets for Random Forest.

Comparison Pair	Feature Set	χ^2 Statistic	p-value	Significant ($\alpha = 0.05$)?
Random Forest vs. SVM	Raw + Indices	0.431	0.512	No
Random Forest vs. XGBoost	Raw + Indices	3.349	0.067	No
Random Forest vs. MLC	Raw + Indices	165.885	< 0.001	Yes
Random Forest: Raw vs. Raw + Indices	Feature-set effect	3.903	0.048	Yes

3.3. Independent Validation Against UAV Orthomosaics

Overall RGB congruence scores between the classified maps and the corresponding UAV orthomosaics are shown in Fig. 6 and ranged from 13.9% to 86.3% across all site–classifier combinations, with broadly overlapping distributions for all four classifiers (mean congruence: MLC = 48.8%, Random Forest = 48.7%, SVM = 45.3%, XGBoost = 44.1%). Unlike the ground-reference-based Overall Accuracy and Kappa results (Section 3.2), the RGB congruence check did not clearly separate MLC from the other three classifiers, and in fact yielded a mean congruence score for MLC on par with Random Forest despite MLC's substantially lower Overall Accuracy. This is consistent with the coarse, category-level nature of the congruence rules described in Section 2.4 (Table 3), in particular the shared rule for Soil, Barren, and Built-up, which is expected to credit pixels that look plausibly "terrestrial" even when the specific class label assigned by a given classifier is incorrect, limiting the check's ability to discriminate classifier quality as finely as the confusion-matrix-based metrics.

4. Discussion

4.1. Classifier Performance in the Context of Previous Studies

The strong, broadly comparable performance of Random Forest, SVM, and XGBoost observed across all five study sites (Section 3.2) is consistent with a growing body of remote sensing literature reporting that modern non-parametric and ensemble classifiers achieve high, often statistically indistinguishable, accuracy on Sentinel-2 land cover classification tasks (Phiri et al., 2020). In a comparable multi-algorithm Sentinel-2 study in a boreal landscape, Abdi (2020) similarly found SVM, XGBoost, and RF clustered closely together in overall accuracy (75.8%, 75.1%, and 73.9%, respectively), with SVM marginally ahead, a ranking broadly echoed in the Raw Bands Only feature set of the present study, where SVM (96.9% mean OA) narrowly outperformed Random Forest (95.5%) and XGBoost (94.9%). Once spectral indices were added, however, Random Forest became the top performer in the present study, a pattern also consistent with mangrove-specific Sentinel-2 classification results reported by Ghorbanian et al. (2021) and Soffianian et al. (2023) on Qeshm Island, and with Seydi et al. (2024), who likewise found ensemble tree-based methods performing competitively with more complex deep-learning architectures once spectrally informative indices were included as inputs.

MLC's markedly weaker performance with raw spectral bands alone, and its correspondingly large gain from added indices (Section 3.2), extends a well-documented limitation of parametric, Gaussian-based classifiers in spectrally complex or heterogeneous landscapes (Phiri et al., 2020). Because MLC assumes that each class's reflectance values follow a multivariate normal distribution with class-specific covariance (Gewali et al., 2018), it is comparatively sensitive to non-linear or overlapping class boundaries, which are likely to arise when only six raw bands are available for a coastal environment spanning distinct mangrove, intertidal, water, and industrial land covers such as around Asalouyeh. Random Forest, SVM (via its non-linear radial basis function kernel), and XGBoost, by contrast, can already partition non-linear decision boundaries directly from the raw bands, which likely explains their

comparatively small and inconsistent response to the addition of spectral indices (Section 3.2). This suggests that spectral index engineering is not a uniformly necessary step for all classifiers, but is disproportionately valuable for simpler, parametric methods such as MLC.

The persistent confusion among Built-up, Barren, and Soil classes across all four classifiers (Section 3.2) is consistent with the broader challenge of separating bright, dry terrestrial cover types in coastal land cover mapping, noted in earlier moderate-resolution studies such as Shalaby and Tateishi's (2007) MLC-based mapping of Egypt's northwestern coast and Kesgin and Nurlu's (2009) classification of Turkey's Candarli Bay coastline. That this confusion persisted even for the three best-performing classifiers in the present study suggests it reflects a genuine spectral limitation of 10 m Sentinel-2 imagery for this class combination, rather than a shortcoming specific to any one algorithm.

4.2. Practical Guidance for Classifier Selection

Taken together, these results offer practical guidance for selecting a classification algorithm in similar coastal and mangrove monitoring contexts. When only the six raw Sentinel-2 reflectance bands are available, or a rapid, minimally pre-processed workflow is preferred, SVM and Random Forest are the most reliable choices, both exceeding 95% mean Overall Accuracy without requiring derived indices; MLC should be avoided under these conditions, since its raw-band accuracy (55.1% mean OA) falls far short of an operationally useful classification. When spectral indices can be computed, as in most standard Sentinel-2 processing pipelines, Random Forest, SVM, and XGBoost remain statistically indistinguishable from one another (Section 3.2), and any of the three can be recommended interchangeably, with the specific choice reasonably guided by secondary considerations such as training time, interpretability, or the analyst's familiarity with a given method. MLC, despite improving substantially with indices, remained significantly less accurate than Random Forest even in this condition (Table 5), and is therefore best reserved for settings where computational simplicity, a small number of training samples, or the interpretability of an explicit parametric model are prioritized over maximum accuracy, and even then only when paired with a carefully engineered feature set rather than raw bands alone.

Because Built-up, Barren, and Soil remained difficult to separate for every classifier tested, none of the four methods can be recommended as a complete, standalone solution where discriminating among these specific classes is the primary monitoring objective, such as tracking industrial expansion around facilities like the Asalouyeh gas complex; complementary data sources (e.g., higher-resolution optical or radar imagery) are likely necessary for that purpose. Finally, the consistency of these patterns across five geographically and environmentally diverse sites, from the relatively undisturbed mangroves of Melgonzeh and Tiyab to the industrially influenced coastline of Asalouyeh, suggests that the relative strengths and weaknesses identified here are likely to generalize reasonably well across the wider Persian Gulf coastal zone, rather than being an artifact of any single site.

4.3. Insights from the Independent UAV-Based Validation

The independent RGB congruence check (Section 3.3) provided a complementary but distinct perspective from the ground-reference-based accuracy metrics. Whereas Overall Accuracy and Kappa

clearly separated MLC from the other three classifiers, the UAV-based congruence scores did not, and even favored MLC slightly. This divergence is informative in itself: it suggests that the coarse, rule-based congruence check described in Section 2.4, while a useful independent sanity check that draws on data entirely outside the training and testing workflow, is not a substitute for, and should not be interpreted as contradicting, a properly labeled ground-reference accuracy assessment. Because three of the eight classes (Soil, Barren, and Built-up) share a single, deliberately broad decision rule (Table 3), and the remaining rules rely on fixed brightness and color thresholds rather than site- or season-calibrated values, the check is better suited to flagging grossly implausible classifications than to finely ranking classifiers that are already broadly accurate. This differs from prior Persian Gulf UAV studies such as Kabiri (2020) and Miraki et al. (2023), which used UAV imagery directly as the primary classification or training data source rather than as an independent post hoc validation layer; the present study's use of UAV orthomosaics purely as an external check, entirely withheld from training and testing, represents, to our knowledge, a novel application of this type of high-resolution reference data in the Persian Gulf coastal classification literature.

4.4. Limitations and Future Research

Several limitations should be considered when interpreting these results. First, accuracy was assessed from a single random 70/30 train–test split per site rather than repeated k-fold cross-validation; although predictions were pooled across all five sites for the confusion matrices and McNemar's tests (Sections 2.3, 3.2), site-level accuracy estimates in Table 4 may be somewhat sensitive to this particular split. Second, classifier hyperparameters (Section 2.3) were fixed rather than individually tuned per site or feature set, so the relative performance of SVM, Random Forest, and XGBoost in particular could shift under a more exhaustive tuning procedure. Third, the RGB congruence rules (Table 3) rely on fixed, manually defined thresholds that were not calibrated against independently verified ground photographs, which likely contributed to the limited discriminative power discussed in Section 4.3; a machine-learned or site-calibrated plausibility model could improve this component in future work. Finally, because all five Sentinel-2 scenes were acquired between February and May across 2020–2023 (Table 1), the relative performance of the four classifiers under different seasonal, tidal, or atmospheric conditions elsewhere in the Persian Gulf remains to be tested. Future studies could extend this comparison to additional seasons, incorporate Sentinel-1 radar data to help resolve the persistent Built-up–Barren–Soil confusion, or apply the same independent, UAV-based validation framework introduced here to other classification comparisons in data-scarce coastal regions.

5. Conclusion

This study set out to determine which of four widely used classification algorithms, Maximum Likelihood Classification, Support Vector Machine, Random Forest, and Extreme Gradient Boosting, is best suited to mapping land cover along a stretch of the Persian Gulf coast that has received far less remote sensing attention than better-studied sites such as Qeshm Island. Rather than pointing to a single universally superior method, the results paint a more nuanced picture: the choice of classifier matters far less than

is sometimes assumed, provided a modern, non-parametric or ensemble-based method is used, but it matters a great deal if a classical parametric approach is chosen without adequate supporting features.

In practice, this means that analysts working in similar coastal and mangrove settings can treat Random Forest, SVM, and XGBoost as broadly interchangeable, reliable workhorses, largely insensitive to whether spectral indices are engineered in addition to the raw satellite bands. Maximum Likelihood Classification, by contrast, is not obsolete, but its viability depends entirely on pairing it with a thoughtfully constructed feature set; used with raw bands alone, it is unsuitable for operational coastal mapping in this environment. This distinction is a useful, transferable lesson for practitioners with limited computational resources or classifier expertise, for whom the simplicity of a parametric method may still be attractive if its known weaknesses are compensated for at the feature-engineering stage.

At the same time, this study makes clear that some land cover distinctions, particularly among built-up, barren, and bare soil surfaces, remain genuinely difficult to resolve from Sentinel-2 imagery regardless of the classification algorithm applied. This is not a shortcoming to be solved by choosing a better classifier, but a reminder that spectral resolution itself imposes a ceiling on what any single-sensor approach can achieve, pointing toward the need for complementary data sources in applications where these specific classes matter most, such as monitoring industrial expansion.

Equally important is what this study demonstrates about validation practice itself. By deliberately withholding independent, high-resolution UAV imagery from both training and testing, and using it purely as an external check, this work shows that ground-reference accuracy and independent physical plausibility can tell different, complementary stories about a classification, and that neither should be treated as a complete picture on its own. This dual-validation approach, though modest in its current implementation, offers a template that other data-scarce, ecologically important coastal regions could adopt to build more trustworthy classification pipelines without requiring extensive dedicated field campaigns.

Ultimately, by bringing together five sites spanning both intact and industrially pressured mangrove and intertidal habitats, this study provides one of the first systematic, multi-site, multi-algorithm evaluations for the Iranian Persian Gulf coast, offering both a practical reference for future monitoring efforts and a foundation for the continued, evidence-based management of a coastline whose ecological value is increasingly at odds with the pace of its industrial development.

Declarations

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This study was supported by the Iran National Science Foundation (INSF) under Project No. 4049590.

Author Contribution

The initial draft of the manuscript was prepared by K.K. The methodology design and programming tasks were jointly conducted by K.K. Final validations and technical checks were performed by K.K. The final review and editing of the manuscript were carried out by K.K.

Data Availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Abdi AM (2020) Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience Remote Sens* 57(1):1–20
2. Afshar-Kaveh N, Nazarali M, Pattiaratchi C (2020) Relationship between the Persian Gulf sea-level fluctuations and meteorological forcing. *J Mar Sci Eng* 8(4):285
3. Bihamta Toosi N, Soffianian AR, Fakheran S, Pourmanafi S, Ginzler C, Waser LT (2020) Land cover classification in mangrove ecosystems based on VHR satellite data and machine learning—An upscaling approach. *Remote Sens* 12(17):2684
4. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
5. Brown K (2004) Increasing classification accuracy of coastal habitats using integrated airborne remote sensing. *EARSeL eProceedings*, 3(1), 34–42
6. Chen Q, Huang M, Wang H (2021) A feature discretization method for classification of high-resolution remote sensing images in coastal areas. *IEEE Trans Geosci Remote Sens* 59(10):8584–8598
7. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM
8. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
9. Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens Environ* 37(1):35–46
10. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297

11. Eskandari S, Jaafari MR, Oliva P, Ghorbanzadeh O, Blaschke T (2020) Mapping land cover and tree canopy cover in Zagros forests of Iran: Application of Sentinel-2, Google Earth, and field data. *Remote Sensing*, 12(12), 1912
12. Feyisa GL, Meilby H, Fensholt R, Proud SR (2014) Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens Environ* 140:23–35
13. Finkl CW (2004) Coastal classification: Systematic approaches to consider in the development of a comprehensive scheme. *J Coastal Res* 20(1):166–213
14. Foody GM (2004) Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Eng Remote Sens* 70(5):627–633
15. Gewali UB, Monteiro ST, Saber E (2018) Machine learning based hyperspectral image analysis: A survey. *arXiv preprint arXiv:1802.08701*
16. Ghorbanian A, Zaghian S, Asiyabi RM, Amani M, Mohammadzadeh A, Jamali S (2021) Mangrove ecosystem mapping using Sentinel-1 and Sentinel-2 satellite images and random forest algorithm in Google Earth Engine. *Remote Sens* 13(13):2565
17. Giang TL, Bui QT, Nguyen TDL, Dang VB, Truong QH, Phan TT, Dang KB (2023) Coastal landscape classification using convolutional neural network and remote sensing data in Vietnam. *J Environ Manage* 335:117537
18. Green EP, Mumby PJ, Edwards AJ, Clark CD (1996) A review of remote sensing for the assessment and management of tropical coastal resources. *Coastal Manage* 24(1):1–40
19. Kabiri K (2020) Mapping coastal ecosystems and features using a low-cost standard drone: Case study, Nayband Bay, Persian Gulf, Iran. *J Coastal Conserv* 24(5):62
20. Kabiri K (2025) Comparative analysis of three key atmospheric correction techniques for bathymetric mapping in nearshore areas with Sentinel-2 Data-case study: Kish Island, Persian Gulf. *Ocean Coastal Res* 73:e25015
21. Kabiri K, Abedi E (2025) Rapid mangrove dieback in the northern Persian Gulf driven by anthropogenic activities and environmental stressors. *Discover Environ* 3(1):22
22. Kesgin B, Nurlu E (2009) Land cover changes on the coastal zone of Candarli Bay, Turkey using remotely sensed data. *Environ Monit Assess* 157(1):89–96
23. Klemas V (2001) Remote sensing of landscape-level coastal environmental indicators. *Environ Manage* 27(1):47–57
24. Klemas V (2011) Remote sensing techniques for studying coastal ecosystems: An overview. *J Coastal Res* 27(1):2–17
25. Lacaux JP, Tourre YM, Vignolles C, Ndione JA, Lafaye M (2007) Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal. *Remote Sens Environ* 106(1):66–74
26. Légaré B, Bélanger S, Singh RK, Bernatchez P, Cusson M (2022) Remote sensing of coastal vegetation phenology in a cold temperate intertidal system: Implications for classification of coastal

- habitats. *Remote Sens* 14(13):3000
27. Liu P, Wang C, Ye M, Han R (2024) Coastal zone classification based on U-net and remote sensing. *Appl Sci* 14(16):7050
 28. Louchard EM, Reid RP, Stephens FC, Davis CO, Leathers RA, Valerie TD (2003) Optical remote sensing of benthic habitats and bathymetry in coastal environments at Lee Stocking Island, Bahamas: A comparative spectral classification approach. *Limnol Oceanogr* 48(1part2):511–521
 29. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157
 30. Miraki M, Sohrabi H, Immitzer M (2023) Tree species mapping in mangrove ecosystems using UAV- RGB imagery and object-based image classification. *J Indian Soc Remote Sens* 51(10):2095–2103
 31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830
 32. Phiri D, Simwanda M, Salekin S, Nyirenda VR, Murayama Y, Ranagalage M (2020) Sentinel-2 data for land cover/use mapping: A review. *Remote Sens* 12(14):2291
 33. Rouse JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication* 351:309–317
 34. Seydi ST, Ahmadi SA, Ghorbanian A, Amani M (2024) Land cover mapping in a mangrove ecosystem using hybrid selective kernel-based convolutional neural networks and multi-temporal Sentinel-2 imagery. *Remote Sens* 16(15):2849
 35. Shalaby A, Tateishi R (2007) Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the northwestern coastal zone of Egypt. *Appl Geogr* 27(1):28–41
 36. Soffianian AR, Toosi B, Asgarian N, Regnauld A, Fakheran H, S., Waser LT (2023) Evaluating resampled and fused Sentinel-2 data and machine-learning algorithms for mangrove mapping in the northern coast of Qeshm island, Iran. *Nat Conserv* 52:1–22
 37. Vanhellemont Q, Ruddick K (2018) Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. *Remote Sens Environ* 216:586–597
 38. Warren MA, Simis SG, Martinez-Vicente V, Poser K, Bresciani M, Alikas K, Ansper A (2019) Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters. *Remote Sens Environ* 225:267–289
 39. Woebbecke DM, Meyer GE, Von Bargen K, Mortensen DA (1995) Color indices for weed identification under various soil, residue, and lighting conditions. *Trans ASAE* 38(1):259–269
 40. Xu H (2006) Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int J Remote Sens* 27(14):3025–3033

Table

Tables 1 is available in the supplementary files section.

Figures

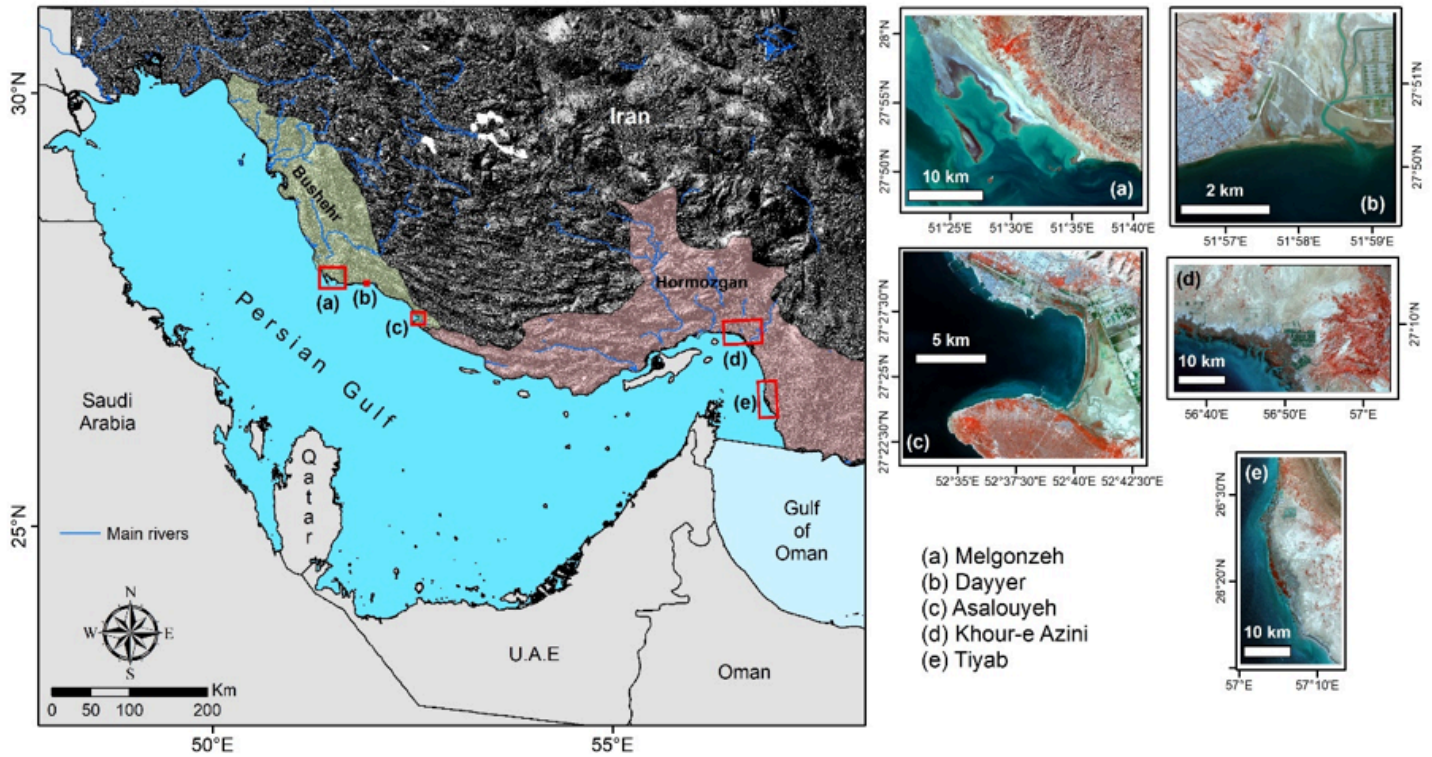


Figure 1

Location of the five study sites along the Persian Gulf coast of southern Iran: (a) Melgonzeh, (b) Dayyer, (c) Asalouyeh, (d) Khour-e Azini, and (e) Tiyyab. Persian Gulf boundaries follow the definition of the International Hydrographic Organization (IHO).

Sentinel-2 LULC Classification Maps Across Coastal Regions (5 Regions × 4 Algorithms)
(Feature Set: Raw Bands + Spectral Indices)

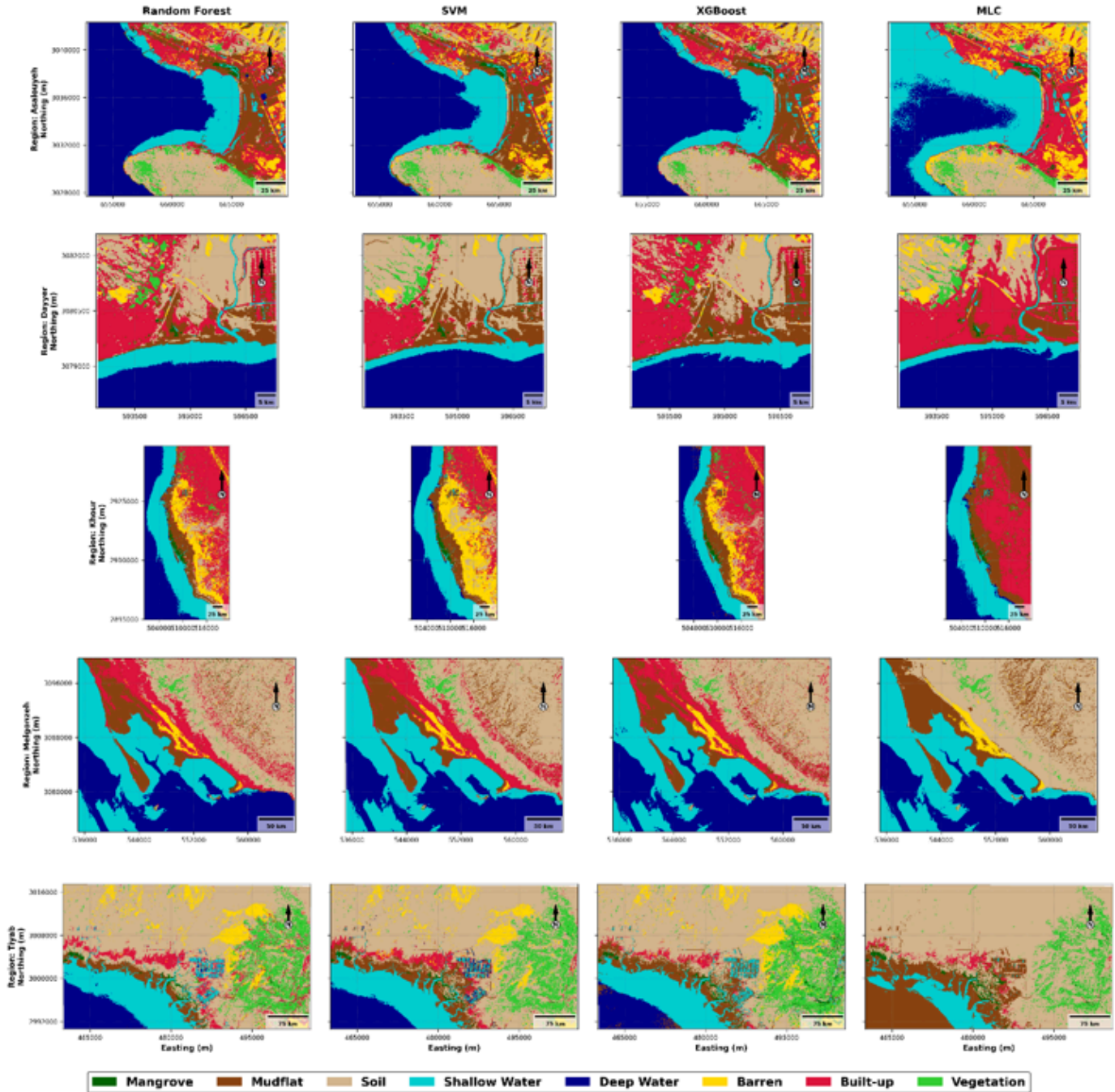


Figure 2

Classified land cover maps for all five study sites (rows) and four classification algorithms (columns), produced using the Raw + Indices feature set. A 3×3 majority filter was applied for visual clarity only (Section 3.1); the accuracy assessment (Section 3.2) and UAV congruence analysis (Section 3.3) used the unfiltered classification outputs. Colors correspond to the eight land cover classes shown in the legend.

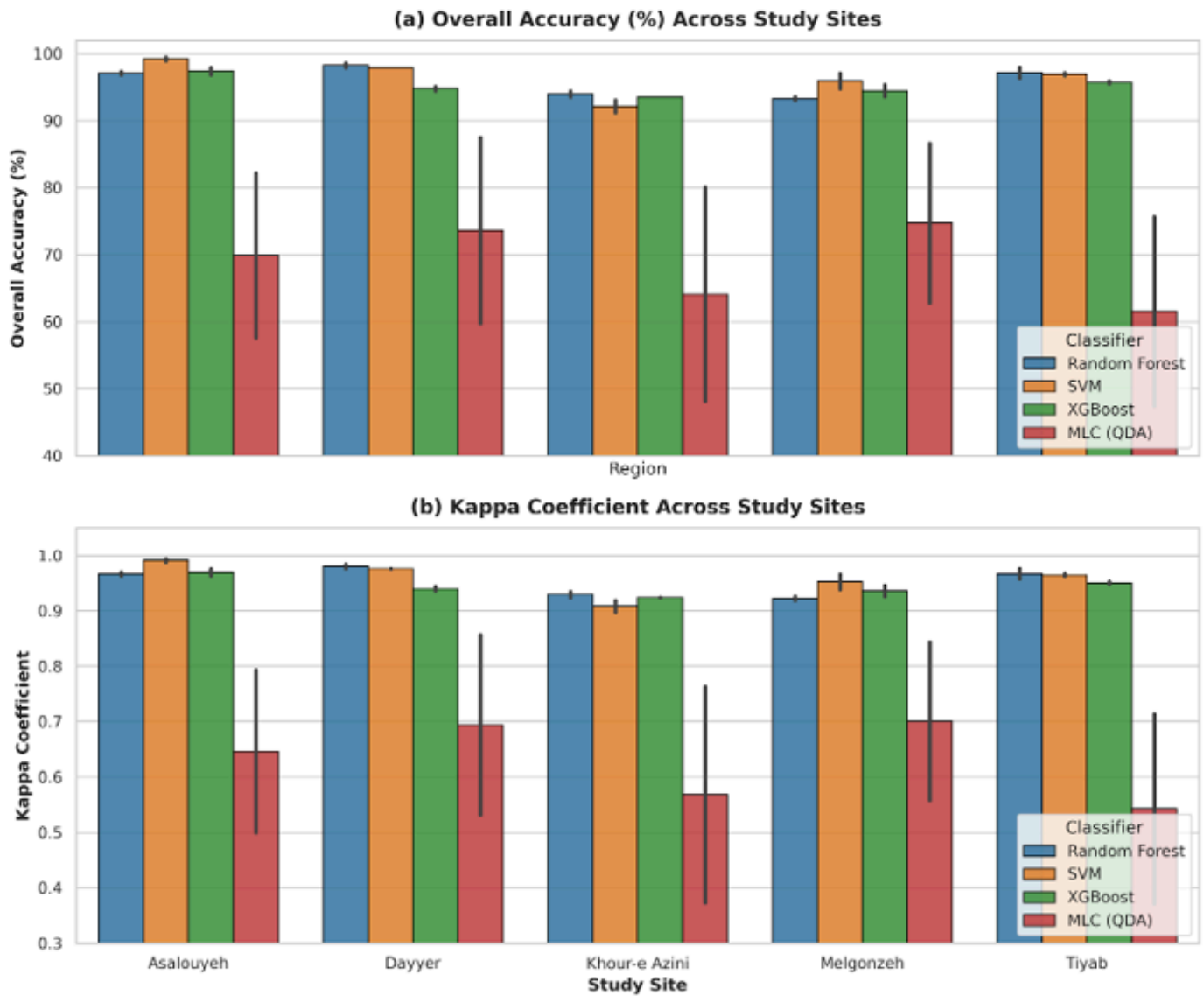


Figure 3

(a) Overall Accuracy (%) and (b) Kappa coefficient for each classifier at each study site. Bar heights show the mean of the Raw + Indices feature sets, and error bars span the values obtained with the two individual feature sets, illustrating how much each classifier's accuracy shifted with the inclusion of spectral indices (Section 3.2).

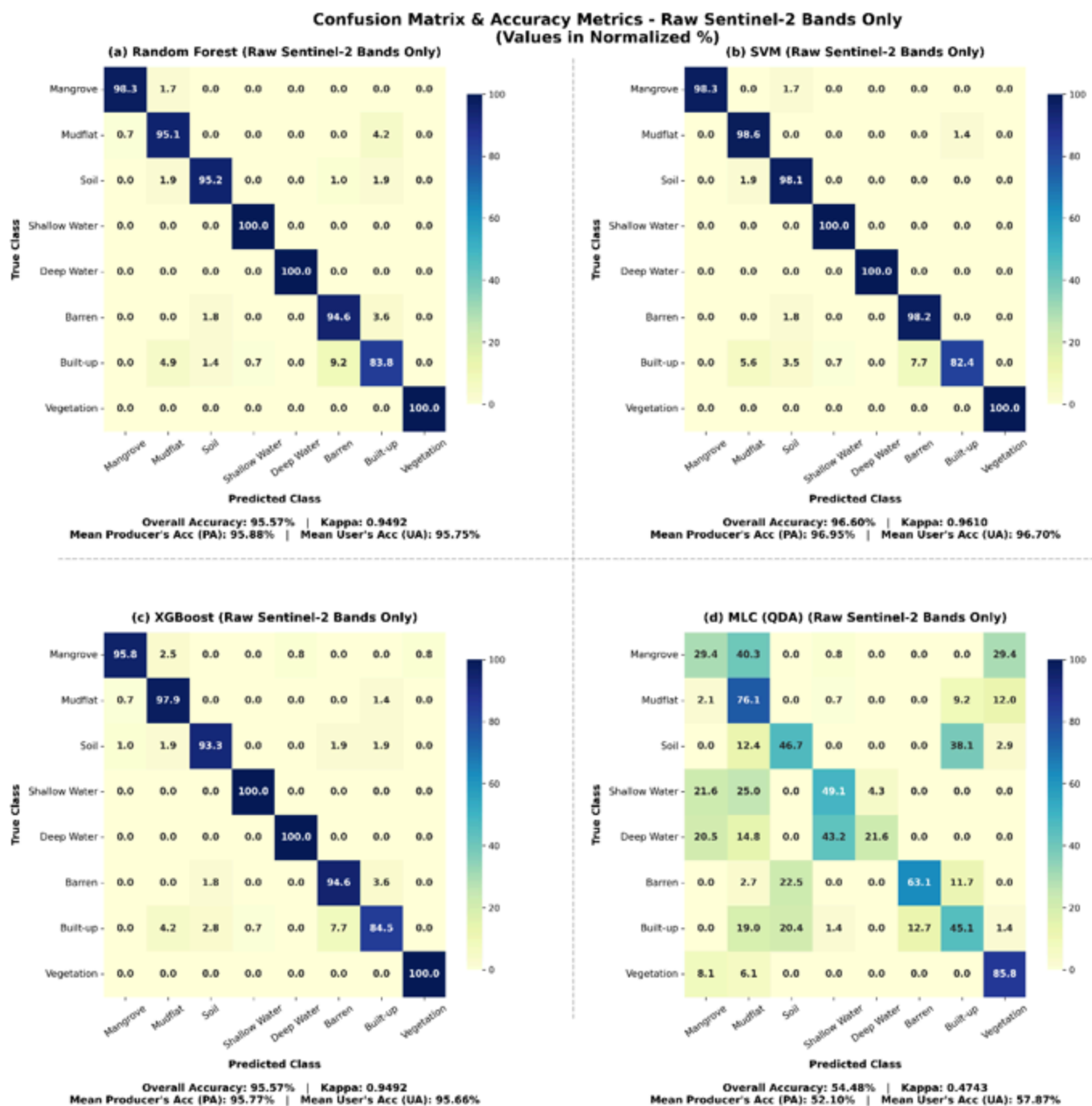


Figure 4

Confusion matrices (row-normalized, %) and accuracy metrics for (a) Random Forest, (b) SVM, (c) XGBoost, and (d) MLC using the Raw Bands Only feature set, with test-set predictions pooled across all five study sites. OA = Overall Accuracy; PA = Producer's Accuracy; UA = User's Accuracy.

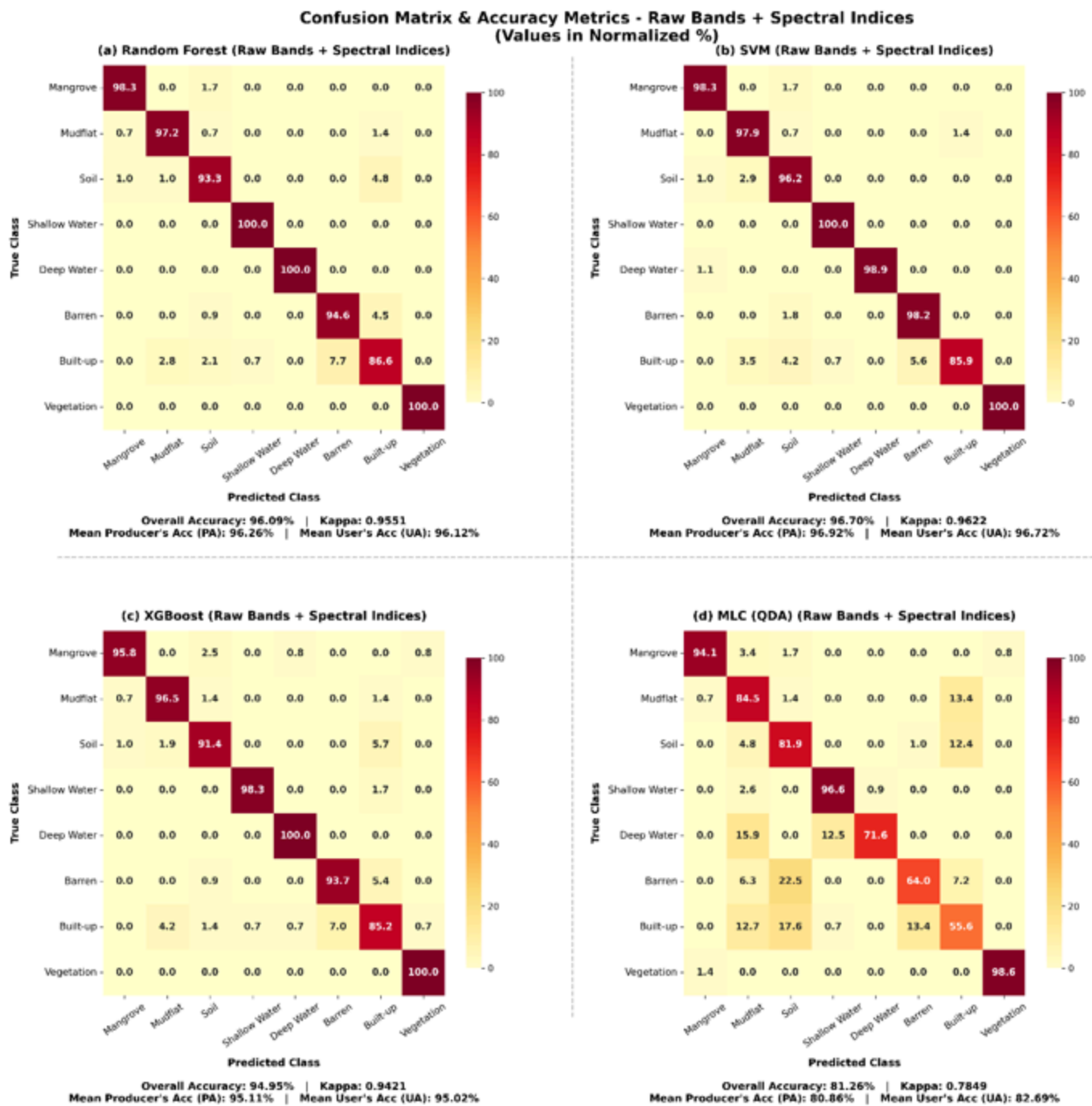


Figure 5

Confusion matrices (row-normalized, %) and accuracy metrics for (a) Random Forest, (b) SVM, (c) XGBoost, and (d) MLC using the Raw + Indices feature set, with test-set predictions pooled across all five study sites. OA = Overall Accuracy; PA = Producer's Accuracy; UA = User's Accuracy.

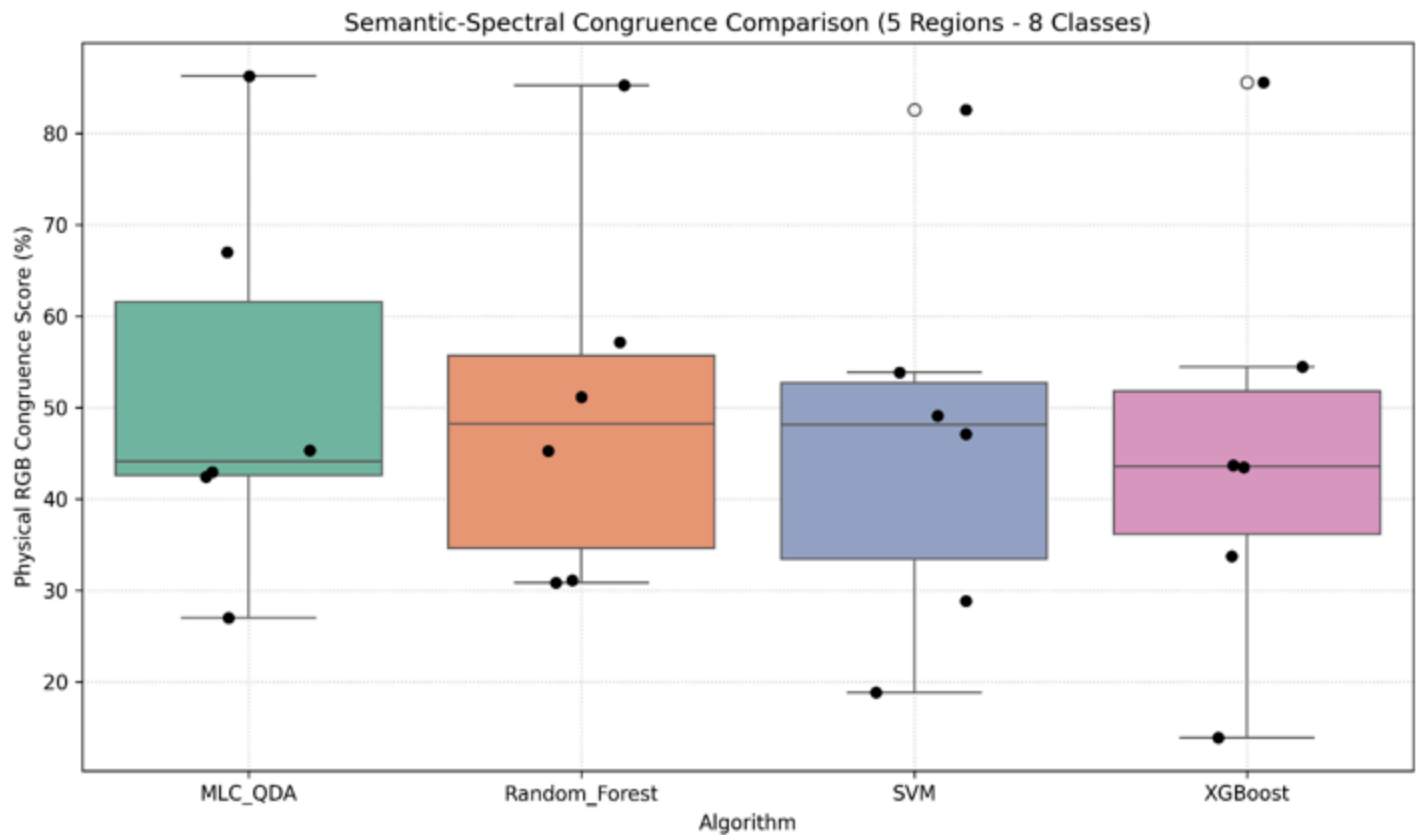


Figure 6

Distribution of overall RGB congruence scores (Section 2.4) between each classifier's output (Raw + Indices feature set) and the corresponding UAV orthomosaic, across all site–classifier combinations. Boxes show the interquartile range and median; points show individual site-level observations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.docx](#)