RESEARCH ARTICLE

How Much Does the Classifier Matter? A Cluster-Aware Comparison of Six Machine Learning Algorithms for Land Use Land Cover Mapping of Taipei, Taiwan, Using Google Earth Engine and Python

Chun-ya Liu[1,*]

[1] Géographie-cités, University of Paris 1 Panthéon-Sorbonne, Aubervilliers 93322, France

Abstract

Comparative studies of machine learning classifiers for Land Use Land Cover (LULC) mapping often recommend one algorithm, yet their validation pixels are usually drawn from a few hand-digitised polygons and are far from independent. This study tested whether such differences survive that dependence. Six classifiers, namely Classification and Regression Tree (CART), Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), a multilayer perceptron and a support vector machine (SVM) with a radial basis function kernel, were applied to four top-of-atmosphere bands of a 2020 annual Sentinel-2 median composite of Taipei City, Taiwan, at 10 m. All 110 reference polygons were kept whole in both the hold-out split and the tuning folds. Pairwise differences were tested with the standard and the cluster-corrected McNemar test, and accuracies and class areas were estimated with the Olofsson stratified estimator on the effective sample size. Hold-out overall accuracy ranged only from 0.927 (CART) to 0.956 (SVM). The 1,829 validation pixels came from 30 polygons and, with a design effect of 71-76, were equivalent to only 24-26 independent observations. Eight of the fifteen pairwise differences were significant under the standard test, but none remained significant after the cluster correction. Mapped areas nevertheless differed markedly between models, from 71.7 to 102.1 km^2 for built-up land and from 10.9 to 32.2 km^2 for swamp; for built-up land, the between-model spread was as large as the sampling uncertainty of a single map. The six classifiers are therefore statistically indistinguishable in accuracy, and the choice among them should rest on computational cost, which differed by a factor of about 330 at inference, and on the stability of the areas they produce.

Keywords: land use land cover, google earth engine, sentinel-2, spatial autocorrelation, clustered McNemar test, design effect, area estimation, Taipei.

Submitted: 21 July 2026
Revised: 11 August 2026
Accepted: 28 September 2026
Published: 26 October 2026

Vol. 2, **No.** 4, 2026.
10.62762/JIAP.2026.490298

***Corresponding author:**
✉ Chun-ya Liu
chunyaliu@hotmail.com

Citation
Liu, C. Y. (2026). How Much Does the Classifier Matter? A Cluster-Aware Comparison of Six Machine Learning Algorithms for Land Use Land Cover Mapping of Taipei, Taiwan, Using Google Earth Engine and Python. *ICCK Journal of Image Analysis and Processing*, 2(4), 252–276.

© 2026 by the Author. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

1 Introduction

Land Use and Land Cover (LULC) datasets underpin urban planning, natural resource inventory, environmental impact assessment and hazard monitoring [1, 2]. As urban expansion intensifies, so does the demand for classification methods that map heterogeneous landscapes accurately and reproducibly [3, 4]. Machine learning (ML) classifiers now dominate the field, and the arrival of cloud platforms such as Google Earth Engine (GEE) [5] has made it straightforward to run several of them over the same imagery and compare the results [6–8].

Comparative studies of this kind are now numerous. Shao and Lunetta [9] compared SVM against CART and neural networks under limited training data; Qian et al. [10] showed that accuracy depends on training sample size for all classifiers tested; Rodriguez-Galiano et al. [11], Ghosh et al. [12] and Zhao et al. [3] reported strong Random Forest performance for land-cover classification, including in urban settings where multi-seasonal imagery and ancillary data are combined; Arpitha et al. [13] confirmed the same ordering for Karnataka, India. The review by Talukdar et al. [14] synthesised more than a hundred such studies. Such comparisons have been conducted across diverse sensors and biomes, including SAR-based classifications of tropical savanna [15], confirming that no single algorithm consistently dominates across all settings. The characteristic result is a table of overall accuracies differing by a few percentage points, followed by a recommendation that the leading algorithm be adopted.

Two methodological problems underlie that recommendation, and neither is usually addressed. The first concerns how the validation sample is built. Training data for LULC classification are almost always collected as hand-digitised polygons, from which pixels are then extracted; a single polygon may yield hundreds of pixels that are spectrally near-identical. If the train/test split is applied to the pixels rather than to the polygons, near-duplicates of the same observation appear on both sides of the split, and the hold-out score reflects memorisation rather than generalisation [16–18]. Even when the split is made correctly at polygon level, the surviving validation pixels remain clustered: 1,829 pixels drawn from 30 polygons are not 1,829 independent tests of the map. Kish's design effect [19] quantifies exactly this loss, yet accuracy figures in the LULC literature are almost universally reported with the nominal count.

The second problem follows from the first. Where a significance test is reported at all, it is McNemar's test on the pixel-level contingency of two maps, which assumes the discordant pixels are independent [20]. Under clustering that assumption inflates the test statistic, and a difference of one or two accuracy points can be declared significant on the strength of sample-size arithmetic alone. Durkalski et al. [21] give a correction for exactly this situation, aggregating discordances within cluster before squaring, but it appears rarely in the remote sensing literature.

Correcting both problems changes the question a comparative study can answer. Instead of asking which classifier is most accurate, one can ask whether the observed accuracy differences are large enough to be distinguished at all given how the reference data were collected-and, if they are not, what else should decide the choice. That is the question this paper takes up. Taipei City provides a demanding test case: a compact basin of roughly 272 km^2 combining dense urban fabric, forested mountains, riverine wetlands and agricultural pockets. The complexity of its hillside land cover transitions has been documented in prior remote sensing work [22, 23], and the heterogeneity of the landscape makes it an informative testbed for classifier comparison. Previous LULC work on Taiwan has concentrated on multi-decadal change at island scale [23, 24] rather than on intra-urban classifier benchmarking.

This study therefore aims to: (1) tune and compare six ML classifiers-CART, Random Forest, XGBoost, LightGBM, a multilayer perceptron and an RBF support vector machine-for LULC mapping of Taipei City on a Sentinel-2 composite, under a partitioning scheme that keeps every training polygon whole; (2) test whether the resulting accuracy differences survive a correction for the clustering of validation pixels within polygons, and quantify the effective sample size that clustering leaves; and (3) determine whether classifier choice affects the class areas derived from the maps more or less than the sampling uncertainty of any one map, since area, not accuracy, is what a planner reads off an LULC product.

2 Study Area

Taipei City, the political, economic and cultural centre of Taiwan, occupies approximately 272 km^2 in the Taipei Basin at the northern tip of the island (Figure 1). Population density is approximately 9,500 persons per km^2 and the city is administratively divided into twelve districts. Taipei City is an enclave of New Taipei City;

together they form the Taipei metropolitan area, home to roughly one-third of Taiwan's population.

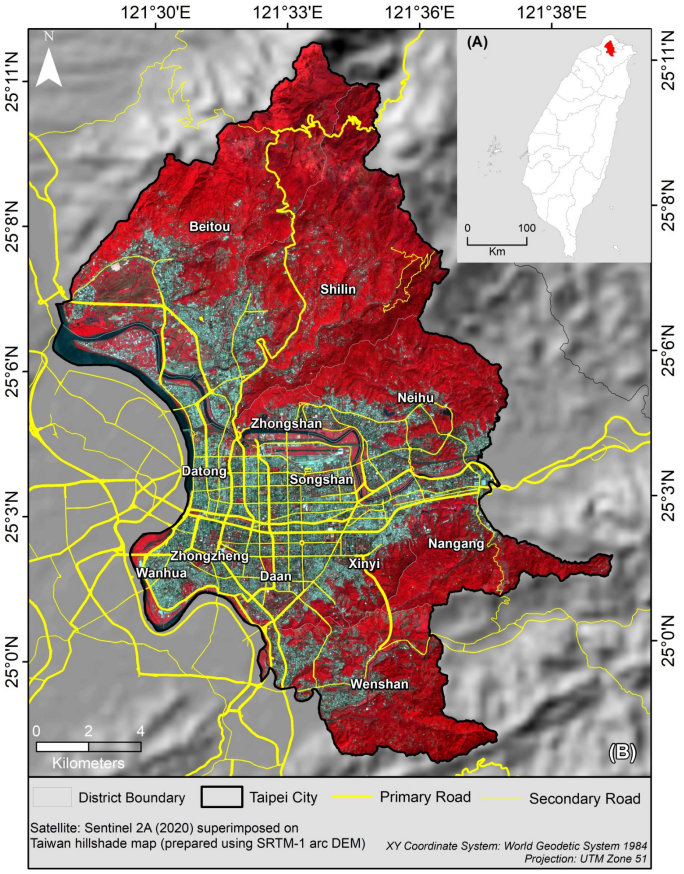


Figure 1. Study area. Taipei City, its twelve administrative districts and its setting in northern Taiwan.

Topography is low-lying in the west and centre and rises into hills to the north, east and south, reaching 1,120 m at Qixing (Chihsin) Mountain, Taiwan's tallest dormant volcano, in the north. Forested hills account for roughly half of the total area and serve as the city's green belt and water catchment. Three major waterways cross the city: the Jingmei River in the south, the Keelung River in the east and the Tamsui River in the west.

Official zoning statistics give public facilities 26.36% of the total area, followed by residential (14.14%), commercial (3.42%) and industrial (1.53%) zones. Non-urban zones comprise conservation land (25.47%), river corridors (6.67%), agricultural land (1.94%) and scenic areas (0.68%). This mosaic, running from dense urban core to mountainous forest within a compact footprint, is what makes Taipei an informative testbed for classifier comparison.

The area actually classified here is 270.44 km^2. The difference from the 272 km^2 administrative figure is the raster footprint: the city boundary is rasterised at 10 m and pixels whose centres fall outside the boundary are dropped. All areas reported below are shares of the 270.44 km^2 classified extent, and are directly comparable with one another.

3 Materials and Methods

The full workflow, from image and vector input through to accuracy and area assessment, is set out in Figure 2.

3.1 Satellite Data and Pre-processing

Sentinel-2 imagery was accessed through the GEE catalog as the `COPERNICUS/S2_HARMONIZED` collection, filtered to the Taipei City boundary and to the calendar year 2020 (1 January to 31 December). Scenes with more than 20% cloud cover, by the `CLOUDY_PIXEL_PERCENTAGE` property, were excluded, cloud and cirrus were masked per pixel using the QA60 band, and the remainder reduced to a per-band median composite. A full-year rather than single-season window was used because Taipei's summer cloud regime leaves too few clear scenes for a stable summer-only median; the median reducer suppresses the residual cloud and shadow that survive the scene-level filter.

This collection is Level-1C, that is, top-of-atmosphere (TOA) rather than atmospherically corrected surface reflectance. It was retained so that all six classifiers are compared on identical inputs. The choice is defensible for a single-year, single-site comparison of this kind, because atmospheric effects are close to common across a composite of one city in one year and therefore shift all six models together rather than favouring any one of them, and because the median reducer over a full year suppresses much of the episodic haze that atmospheric correction targets. It would not be defensible for multi-temporal change detection or for transferring a trained model to another date or site, and it is one reason the absolute accuracies here should be read as conditional rather than as a benchmark.

Four Multispectral Instrument (MSI) bands were retained, all at 10 m native resolution: blue (B2, 490 nm), green (B3, 560 nm), red (B4, 665 nm) and near-infrared (B8, 842 nm) (Table 1). Reflectance was scaled to the 0-1 range. The composite was projected to Universal Transverse Mercator (UTM) zone 51N (WGS 84) at 10 m so that the areas reported below are measured on a metric grid.

One detail of the GEE implementation is worth recording because it is easy to miss and materially

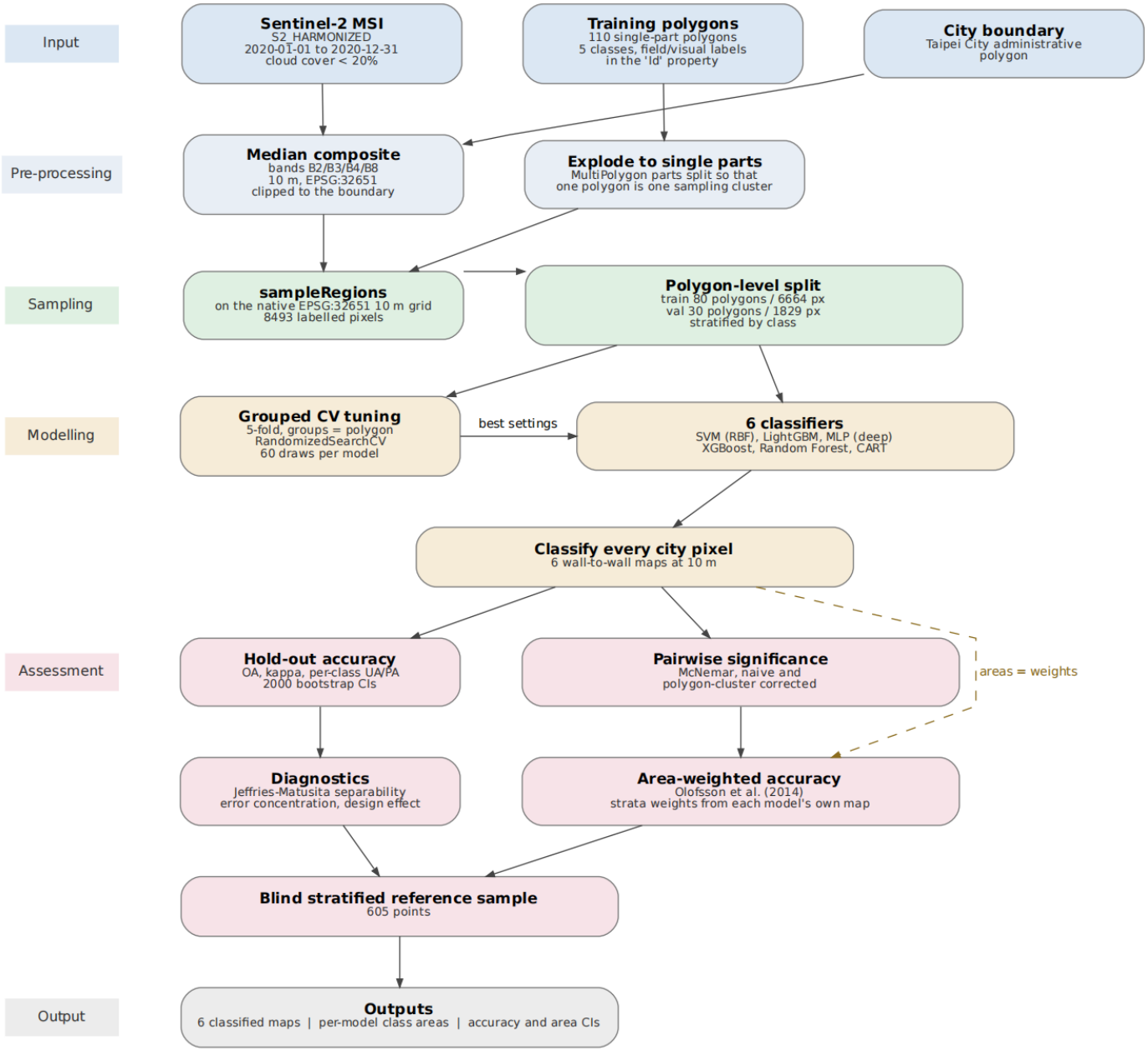


Figure 2. Methodology. Processing runs top to bottom through six stages. Counts inside the boxes are read from the analysis outputs. The dashed feedback path marks the one circularity in the design: the stratum weights W_i of the area-weighted estimator are the mapped class areas, so each model must be assessed against a sample stratified by its own map.

affects the sample. A composite produced by `ee.ImageCollection.median()` carries a default geographic projection at one degree per pixel. Calling `sampleRegions()` on such an image without an explicit `projection` argument therefore samples on the default grid rather than the intended 10 m grid, which in our first run produced 8.8% duplicated rows in the training table. The scale and projection are set explicitly at every sampling call in the released code.

Table 1. Sentinel-2 MSI bands used for classification, from the Level-1C top-of-atmosphere `COPERNICUS/S2_HARMONIZED` collection.

Band	Description	λ (nm)	Res. (m)
B2	Blue	490	10
B3	Green	560	10
B4	Red	665	10
B8	Near-infrared	842	10

3.2 Training Sample Design

Reference polygons for five LULC classes-water, built-up, forest/vegetation, plantation and swamp-were digitised on-screen over high-resolution basemap imagery, guided by field familiarity with the city. Class definitions are as follows: water covers rivers and open water bodies; built-up covers buildings, roads and other impervious surfaces; forest/vegetation covers the hill forest and dense urban greenery; plantation covers cultivated and managed vegetation including urban parks and riparian cultivation; and swamp covers the seasonally inundated margins of the Tamsui and Keelung rivers.

The digitised layer was stored as one feature per class, each a MultiPolygon. Because the unit of spatial dependence is the individual polygon rather than the class, every MultiPolygon was exploded into its constituent parts before any partitioning, giving 110 separate polygons and 8,493 labelled 10 m pixels (Table 2; Figure 3(a) maps their distribution across the city).

Table 2. Labelled pixels per class after the polygon-grouped split (80 training and 30 validation polygons, nominally 70/30). The split is applied to the 110 reference polygons, not to the pixels, so no polygon contributes to both columns.

Class	Training	Validation	Total
Water	1,551	153	1,704
Built-up	1,577	440	2,017
Forest/Vegetation	2,383	970	3,353
Plantation	775	150	925
Swamp	378	116	494
Total	6,664	1,829	8,493

Polygons were then split approximately 70/30 (80 and 30 polygons) into training and validation sets, stratified by class, with the split applied to *polygons* and not to pixels. Every pixel of a polygon therefore falls entirely on one side of the split. This is the most consequential methodological choice in the study. Figure 4(a) measures why it is necessary. Pixels drawn from within one digitised polygon are near-duplicates of one another: over pairs of labelled pixels of the same class, the median Euclidean distance in the four-band reflectance space is 3.5 times larger when the two pixels come from different polygons than when they come from the same one. A pixel-level split therefore places copies of the same observation on both sides of the divide, and the resulting hold-out score is optimistically biased by an amount that cannot be estimated after the fact [16–18]. The split yielded 80 training polygons (6,664 pixels) and 30 validation polygons (1,829 pixels).

3.3 Classification Algorithms

Six classifiers were applied to the same four-band feature vector, the same training partition and the same tuning budget. They span the four families in common use for LULC work (a single tree, three tree ensembles, a margin method and a neural network), so that any difference between the resulting maps is attributable to the decision rule alone.

3.3.1 *Classification and Regression Tree*

CART [25] recursively partitions the feature space with axis-aligned splits chosen to minimise node impurity. For a node t with class proportions $p(i \mid t)$ the Gini index is

$$G(t) = 1 - \sum_{i=1}^{k} p(i \mid t)^2, \tag{1}$$

and the split maximising the impurity decrease $\Delta G = G(t) - p_L G(t_L) - p_R G(t_R)$ is selected at each step. The tree is post-pruned by minimal cost-complexity pruning with parameter α [25].

3.3.2 *Random Forest*

Random Forest [26, 27] grows an ensemble of trees, each on a bootstrap resample of the training set and each split restricted to a random subset of features, and assigns the class by majority vote:

$$\hat{y}(\mathbf{x}) = \arg\max_{c} \sum_{b=1}^{B} \mathbb{I}[h_b(\mathbf{x}) = c], \tag{2}$$

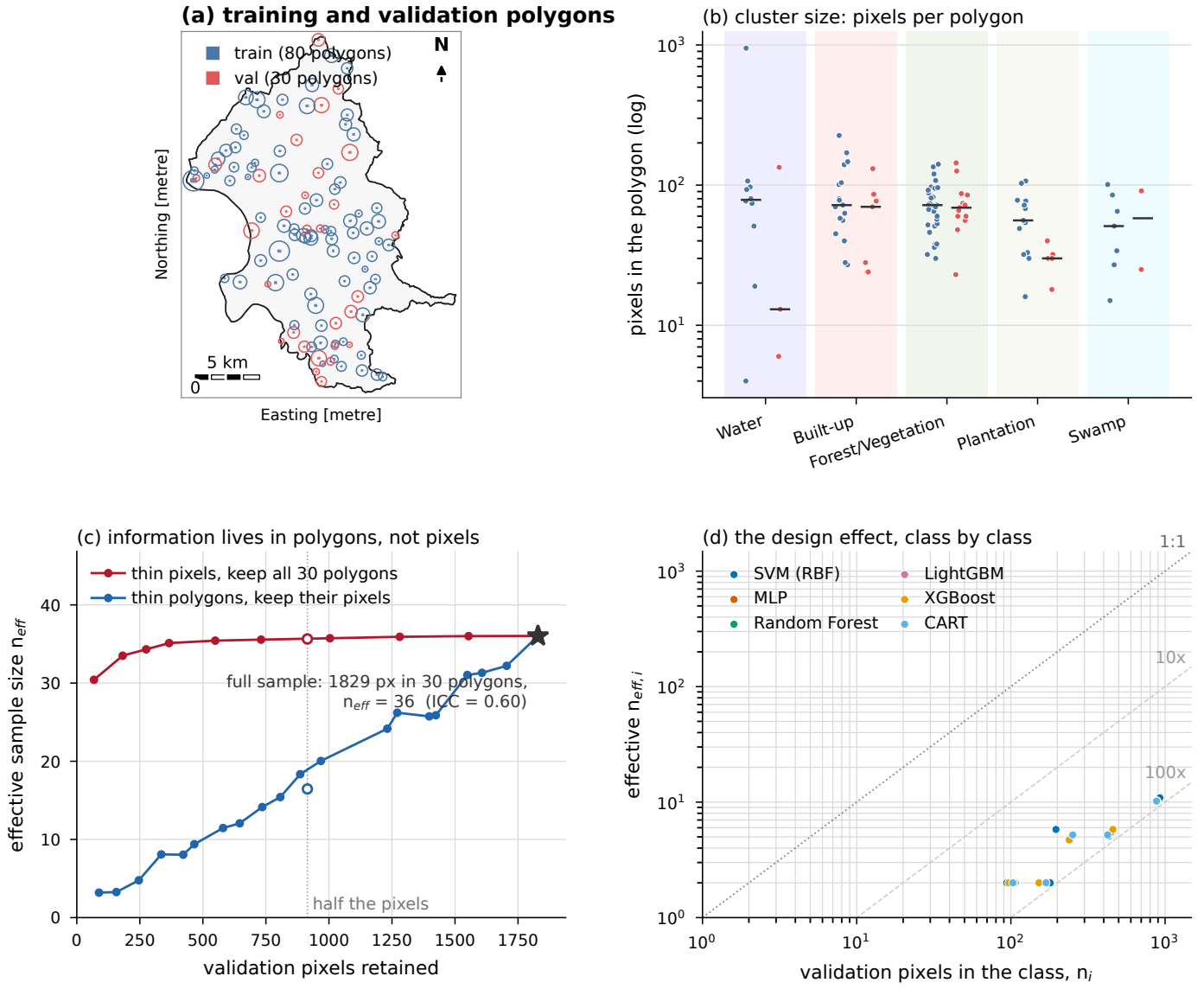
where h_b is the b-th tree. Randomising both rows and columns decorrelates the trees and reduces the variance of the aggregate relative to any single tree.

3.3.3 *Gradient-Boosted Trees*

XGBoost [28] and LightGBM [29] both fit trees in sequence, each new tree fitted to the gradient of the loss of the current ensemble. Writing $\mathcal{L}$ for the loss and Ω for a complexity penalty, the objective at boosting round m is

$$\mathcal{J}^{(m)} = \sum_{n} \mathcal{L}\Big(y_n,\, \hat{y}_n^{(m-1)} + f_m(\mathbf{x}_n)\Big) + \Omega(f_m). \tag{3}$$

The two differ mainly in how they grow each tree: XGBoost grows trees level-wise (depth-wise) by default, whereas LightGBM grows them leaf-wise and adds histogram-based split finding with optional gradient-based one-side sampling (GOSS) [29]. This matters for training cost more than for the fitted decision surface.



panel (c) uses SVM (RBF), the highest-OA model; the other five give the same shape

Figure 3. Sampling design and effective sample size. (a) The 110 training polygons, 80 training and 30 validation; ring size is proportional to pixel count. (b) Pixels per polygon by class and split, log scale. (c) The cost of clustering: thinning pixels within polygons (red) against removing whole polygons (blue). (d) The design effect class by class and model by model.

3.3.4 Support Vector Machine

SVM [30, 31] seeks the hyperplane of maximum margin in a space induced by a kernel. With the radial basis function kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \qquad (4)$$

the decision function is $f(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$, with the α_i obtained from the soft-margin dual problem under the box constraint $0 \le \alpha_i \le C$. Features were standardised before fitting, as the RBF kernel is not scale invariant [32].

3.3.5 Multilayer Perceptron

The MLP is a feedforward network with one hidden layer of rectified linear units (ReLU) and a softmax output, trained by Adam on the cross-entropy loss with L_2 weight decay α. As with the SVM, features were standardised first.

3.3.6 Hyperparameter Tuning

For the SVM, C and γ were the primary tuning targets; the sensitivity of RBF-SVM performance to these parameters in remote sensing applications has been demonstrated in operational mapping contexts [32, 33]. Each model was tuned by randomised search

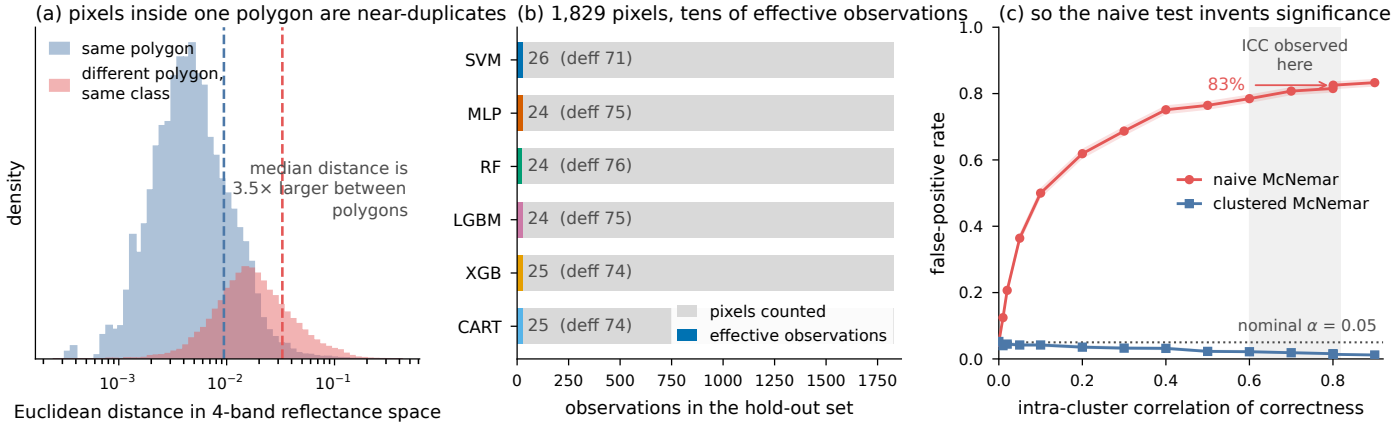


Figure 4. Why the model comparison has to be cluster-aware, in three measured steps. (a) Pixels inside one digitised polygon are near-duplicates of one another: for pairs of pixels of the same class, the median distance in the four-band reflectance space is 3.5 times larger between polygons than within one. (b) The 1,829 validation pixels are therefore worth only 24 to 26 effective observations, a design effect of 71 to 76. (c) A test that assumes independence over-rejects accordingly. Under a null in which two classifiers have identical expected accuracy, the false-positive rate of the naive McNemar test climbs with the intra-cluster correlation of correctness and reaches 82.5% at the value observed here, while the cluster-corrected test of Durkalski et al. [21] stays at or below the nominal 5%. Both tests are calibrated at zero correlation, which is what validates the two implementations. Four thousand null replicates per point; bands are 95% binomial intervals.

over 60 candidate configurations with five-fold cross-validation. Critically, the folds were grouped by polygon, using the same grouping variable as the outer split, so that no polygon appeared in both a fitting fold and a scoring fold. Tuning on ungrouped folds would select the configuration that best memorises polygon identity, which is not the configuration that generalises. Selected values are given in Table A1.

3.4 Accuracy Assessment

Accuracy was assessed at three levels of rigour, and the differences between them are a substantive result of this study rather than a technical preliminary.

3.4.1 *Hold-out accuracy*

Overall accuracy (OA) and Cohen's κ [34] were computed on the 1,829 held-out pixels, with 95% confidence intervals from 2,000 bootstrap replicates. These figures are reported for comparability with the existing literature, in which they are the standard currency, and are not the basis for any inference here.

3.4.2 *Effective sample size*

The 1,829 validation pixels come from only 30 polygons. Treating them as independent overstates the information they carry by the design effect [19]

$$\mathrm{deff} = 1 + (\bar{m}_w - 1)\,\rho, \qquad \bar{m}_w = \frac{\sum_k m_k^2}{\sum_k m_k}, \tag{5}$$

where m_k is the number of validation pixels in polygon k and $\bar{m}_w$ is the size-weighted mean cluster size. $\bar{m}_w$ equals the ordinary mean when all polygons are the same size and exceeds it otherwise; here the validation polygons hold 6 to 144 pixels, and $\bar{m}_w = 83.9$ against an arithmetic mean of 61.0. ρ is the intra-cluster correlation (ICC) of the per-pixel correctness indicator, estimated by one-way analysis of variance with the adjusted cluster size $m_0 = (n - \sum_k m_k^2/n)/(k-1)$ for unequal clusters. Because the Olofsson estimator is stratified by map class, deff_i is computed separately within each map stratum i, and $n_{\mathrm{eff},i} = n_i/\mathrm{deff}_i$ (floored at 2). Where a stratum contains no misclassified pixels, ρ is undefined and its limiting value $\rho = 1$ is used, giving $n_{\mathrm{eff},i} = n_i^2/\sum_k m_{ik}^2$. Every variance below is computed on $n_{\mathrm{eff},i}$ rather than on n_i. The effect is large here: summed over strata, the 1,829 validation pixels carry the information of only 24 to 26 independent observations, depending on the model (Figure 4(b)). This is fewer than the 30 polygons, because the polygons are unequal in size and the largest of them dominate the variance.

3.4.3 *Area-weighted accuracy and area estimation*

Because the validation sample is not proportional to the mapped extent of each class, the raw confusion matrix is a biased estimator of map accuracy. The stratified estimator of Olofsson et al. [35] and Stehman [36] reweights each cell by the mapped area share W_i of stratum i:

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i\cdot}}, \qquad \widehat{\mathrm{OA}} = \sum_i \hat{p}_{ii}. \tag{6}$$

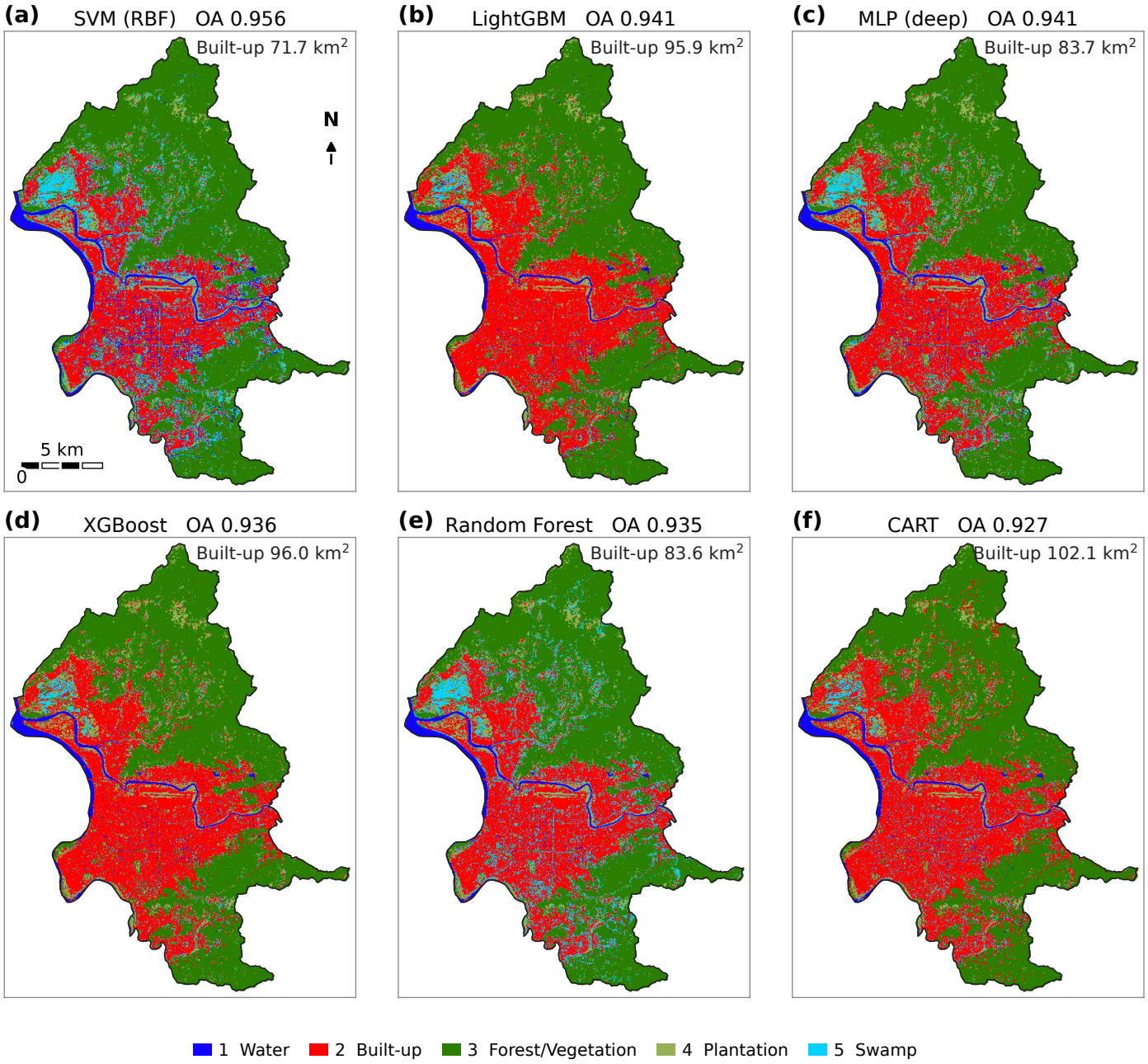


Figure 5. Six-panel comparison of the 2020 LULC maps of Taipei City produced by the six classifiers at 10 m. Panels are ordered by hold-out overall accuracy, which is printed with each model name; mapped built-up area is annotated at the lower right of each panel.

The area of class j follows as $\hat{A}_j = A_{\mathrm{tot}} \sum_i \hat{p}_{ij}$, with standard error

$$S(\hat{A}_j) = A_{\mathrm{tot}} \sqrt{\sum_i W_i^2 \, \frac{\hat{p}_{ij}/W_i \, (1 - \hat{p}_{ij}/W_i)}{n_{\mathrm{eff},i} - 1}}. \tag{7}$$

Substituting $n_{\mathrm{eff},i}$ for $n_{i\cdot}$ in Eq. 7 is the one departure from the published formulation, and it keeps the intervals consistent with the polygon design; without it they would be roughly 8.4 to 8.7 times too narrow ($\sqrt{\mathrm{deff}}$). Note that the weights W_i are the class areas of the map being assessed, so each model is necessarily evaluated against a sample stratified by its own map-the feedback path marked in Figure 2.

3.4.4 *Testing differences between models*

Pairs of models were compared with McNemar's test on the discordant pixels. In its standard form, with b and c the counts of pixels correct under exactly one of the two models,

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \tag{8}$$

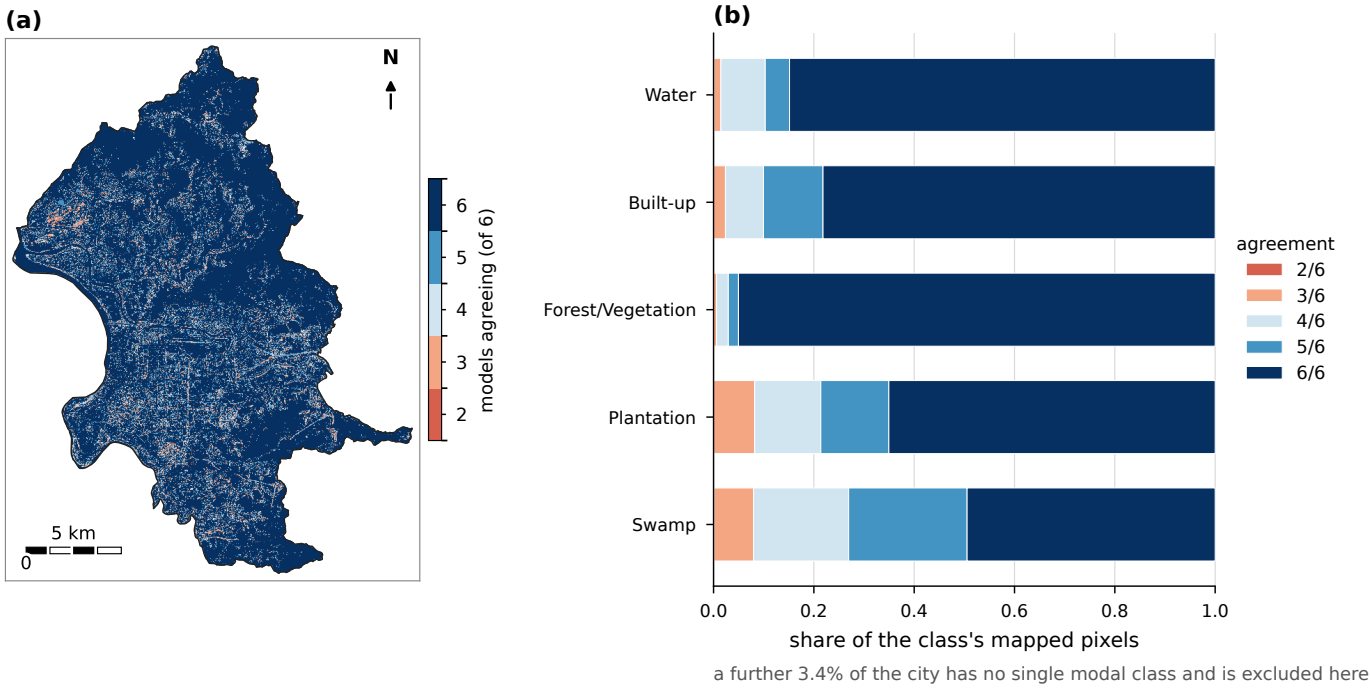


Figure 6. Spatial structure of disagreement between models. Left: for every pixel, the number of the six classifiers assigning it the modal class. Right: for each modal class, the share of its pixels at each agreement level.

which assumes the discordances are independent. Under the polygon design they are not, and the correction of Durkalski et al. [21] was applied: discordant outcomes are summed within polygon before squaring, so that a run of errors inside one polygon counts once rather than many times. Both versions were computed for all fifteen pairs and Holm-adjusted for multiplicity within each version [37].

The validity of the correction under this design was checked by simulation rather than assumed. Null data were generated on the observed design-the 30 validation polygons at their observed sizes, two classifiers of identical expected accuracy, and a per-polygon error rate drawn independently for each model from a Beta distribution tuned to a given intra-cluster correlation ρ-so that every rejection is a false positive by construction. Four thousand replicates were run at each of fourteen values of ρ (Figure 4(c)). At $\rho = 0$, where the pixels really are independent, both tests are calibrated: the naive test rejects at 5.0% and the clustered test at 5.2%, against a nominal α of 0.05. At the $\rho = 0.80$ observed here-the median over the six models, which span 0.60 to 0.82-the naive test rejects at 82.5% while the clustered test rejects at 1.4%. The correction is therefore conservative on this design, which is the direction that protects a negative finding.

3.4.5 Class separability

The Jeffries-Matusita distance was computed for all ten class pairs in the four-band feature space on the training pixels. For two classes assumed Gaussian with means $\boldsymbol{\mu}_i$ and covariances Σ_i, the Bhattacharyya distance is

$$B = \tfrac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\top}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \tfrac{1}{2}\ln\frac{|\Sigma|}{\sqrt{|\Sigma_1|\,|\Sigma_2|}}, \quad (9)$$

with $\Sigma = (\Sigma_1 + \Sigma_2)/2$, and $\mathrm{JM} = 2\left(1 - e^{-B}\right)$, bounded above by 2. Pairs below approximately 1.9 are conventionally read as not reliably separable. This diagnostic is a property of the band set alone and is independent of any classifier, which makes it the natural place to look when all six models fail in the same way.

4 Results

Results are presented in the order in which the argument is built: the six maps and how similar they look; their accuracies and the intervals around them; whether the differences between those accuracies are real; how far the class areas move between models; and where in the city and in the feature space the surviving errors sit.

4.1 Classified Maps

All six classifiers produced a complete LULC map of the 270.44 km^2 study extent at 10 m (Figure 5).

At the scale of the city the six are hard to tell apart: the forested Beitou, Shilin, Neihu and Wenshan hills, the built-up basin floor and the three river corridors appear in every map, and the six agree pixel-for-pixel on 81% of the city.

That agreement is not evenly distributed. Figure 6 shows 94% of pixels mapped as forest/vegetation are unanimous across all six models, against only 50% of those mapped as swamp and 64% of those mapped as plantation. The disagreement is spatially organised rather than scattered: it forms narrow bands along class boundaries and concentrates in the basin, which is the pattern expected of a genuine ambiguity in the class definitions rather than of independent classifier noise.

Figure 7 shows the most accurate map, the RBF support vector machine, at full page with district boundaries and three enlarged windows. Two features of the map deserve comment. In the Beitou and Shilin foothills (window A) the forest-plantation boundary follows the terrain break closely, which is the behaviour one wants. In the Zhongshan-Datong-Songshan core (window B) the wide boulevards of the basin floor are assigned to water and swamp rather than to built-up. This is the one systematic failure that survives in the best model, and Section 4.5 shows it to be a property of the four-band feature set rather than of the SVM.

4.2 Accuracy and Its Uncertainty

Hold-out overall accuracy spans 2.96 percentage points, from 0.9267 for CART to 0.9563 for the RBF SVM, with κ from 0.8900 to 0.9332 (Table 3). The ordering-SVM, then LightGBM and the MLP tied, then XGBoost, Random Forest and CART-is the ordering a conventional comparative study would report, and the bootstrap intervals around it are narrow: the SVM interval [0.9464, 0.9650] does not overlap the CART interval [0.9142, 0.9382]. Cross-validated accuracy on the grouped folds is uniformly higher than hold-out accuracy, by between 2.5 and 4.8 points, which is the expected direction: the configuration of each model was selected by maximising this same cross-validated score over 60 candidates, so the score is optimistically biased, and the scoring folds are drawn from the same pool of polygons used for tuning.

Those bootstrap intervals, however, resample the 1,829 validation pixels as if they were independent draws, which the polygon design does not license. The intra-cluster correlation of the correctness indicator, estimated within each map stratum across the validation polygons, gives overall design effects (n/n_{eff}) between 70.8 and 75.5 depending on the model. The 1,829 pixels are therefore worth between 24.2 and 25.8 independent observations (Table 4).

Recomputing the accuracies on that effective sample, and reweighting the confusion matrices by mapped area after Olofsson et al. [35], gives a different result. Area-weighted overall accuracy ranges from 0.9484 (XGBoost) to 0.9701 (SVM), but the standard errors are between 0.0320 and 0.0426, so every interval covers every other model's point estimate. Figure 8 sets the two panels side by side: the apparent separation between the six models in the hold-out panels is finer than the precision the sampling design can support.

The per-class picture (Table A2, Figure 9) locates the weakness. User's accuracy for forest/vegetation is at least 0.998 in every model and for built-up at least 0.950. Plantation is the weak class throughout: user's accuracy runs from 0.595 (CART) to 0.755 (SVM), meaning that between a quarter and two-fifths of the pixels mapped as plantation belong to another class, almost always forest/vegetation. Producer's accuracy for plantation, by contrast, is 0.98 or better in every model-the classifiers find the plantation, but over-assign to it.

4.3 Significance of Between-Model Differences

The fifteen pairwise McNemar tests are reported in Table A3 and plotted in Figure 10. Treating the validation pixels as independent, eight of the fifteen differences are significant at $\alpha = 0.05$ after Holm adjustment, including every comparison involving CART except CART-XGBoost and every comparison involving the SVM. The smallest naive p-value, for CART against the SVM, is 2.9×10^{-8}.

Applying the cluster correction of Durkalski et al. [21], which aggregates discordant outcomes within polygon before squaring, *none* of the fifteen differences remains significant. The smallest adjusted clustered p-value across all fifteen pairs is 0.293, for CART against the MLP; the other fourteen are all 1.000 after adjustment. The reason is visible in the discordant-polygon counts: the 78 pixels on which CART and the SVM disagree (66 in the SVM's favour) come from only ten polygons, so the effective evidence for a difference rests on ten clusters rather than 78 pixels.

The two columns thus lead to different conclusions: the naive test would support recommending one algorithm, whereas the clustered test cannot distinguish any of the six.

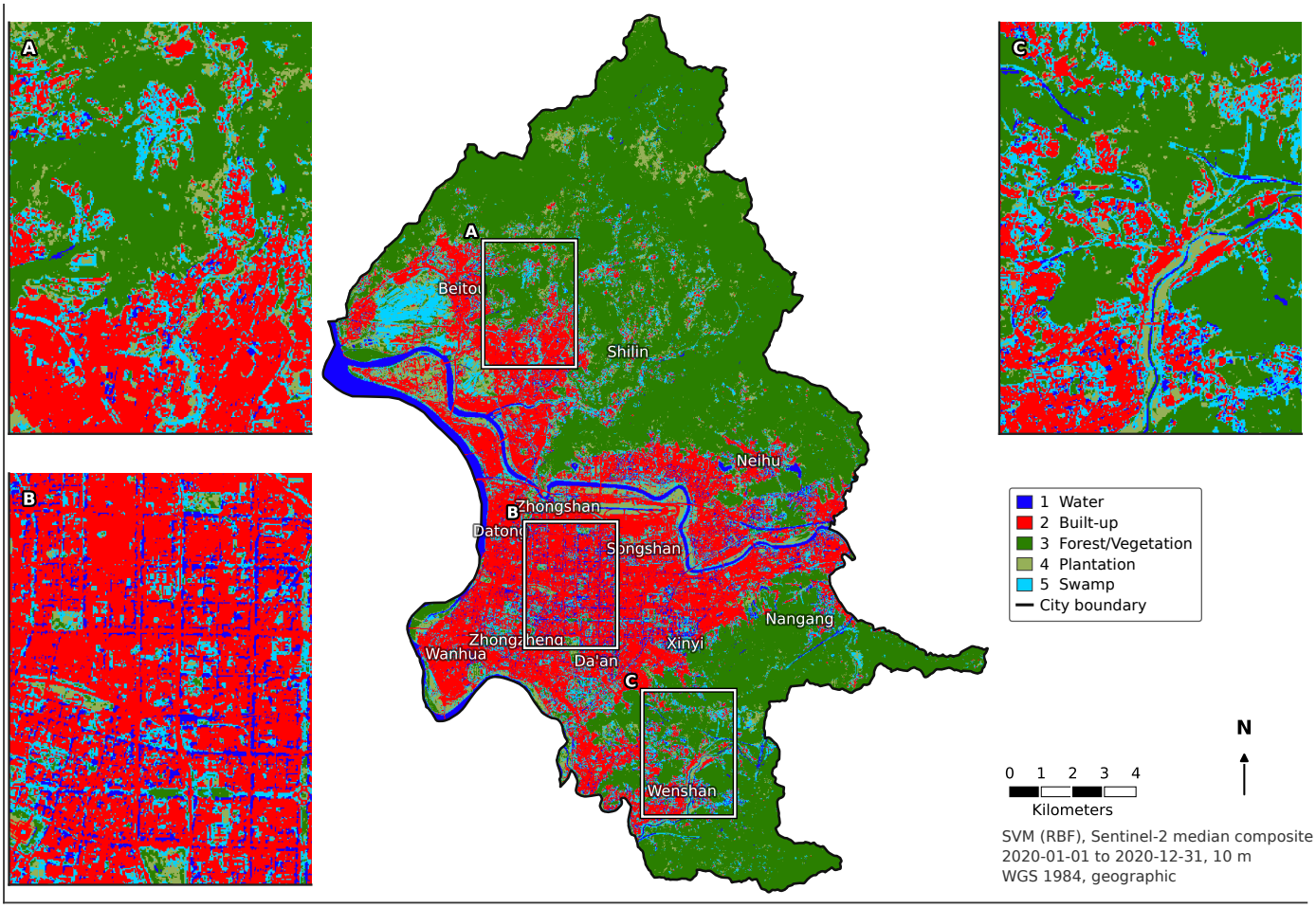


Figure 7. LULC of Taipei City in 2020 from the most accurate classifier, SVM with an RBF kernel. The twelve districts are named and outlined in white. Three windows are enlarged in the margins: (A) the Beitou and Shilin foothills; (B) the Zhongshan-Datong-Songshan core; (C) the Wenshan valley.

Table 3. Hold-out performance of the six tuned classifiers, ordered by overall accuracy. Cross-validated accuracy is the mean over five polygon-grouped folds of the training partition. Confidence intervals are 2000-replicate bootstrap percentile intervals over the 1,829 validation pixels and, because that resampling treats the pixels as independent, are reported here only for comparison with Table 4. Tuning time is wall-clock time for the full randomised search; inference time is for the complete validation set.

Model	CV acc.	OA	95% CI	κ	95% CI	Tuning (s)	Inference (ms)
SVM (RBF)	0.9808	0.9563	[0.9464, 0.9650]	0.9332	[0.9185, 0.9467]	6.4	106.23
LightGBM	0.9681	0.9410	[0.9306, 0.9513]	0.9111	[0.8948, 0.9264]	1,124.5	40.24
MLP	0.9778	0.9410	[0.9305, 0.9513]	0.9111	[0.8954, 0.9268]	98.9	1.08
XGBoost	0.9657	0.9360	[0.9251, 0.9470]	0.9035	[0.8870, 0.9197]	19.1	2.86
Random Forest	0.9788	0.9355	[0.9240, 0.9459]	0.9030	[0.8860, 0.9188]	54.6	14.37
CART	0.9744	0.9267	[0.9142, 0.9382]	0.8900	[0.8720, 0.9073]	3.0	0.32

4.4 Between-Model Differences in Class Area

Statistical indistinguishability in accuracy does not imply that the six maps are interchangeable. Table 5 gives the mapped area of each class in each map. Built-up land ranges from 71.68 km^2 (SVM) to 102.06 km^2 (CART), a spread of 30.38 km^2, or 1.42 times the smallest estimate-more than 11% of the entire city. Swamp ranges from 10.88 km^2 (LightGBM) to 32.16 km^2 (Random Forest), a factor of 2.96. Water ranges over a factor of 1.96. Only forest/vegetation is stable, spanning 126.64 to 133.34 km^2, a spread of 5%.

The comparison that matters for a user of these maps is between that between-model spread and the sampling uncertainty of any one model's area estimate. Table 6 and Figure 11 make it directly. For built-up land, half the between-model range is 15.19 km^2, the same as the

Table 4. Area-weighted overall accuracy after Olofsson et al. [35], with the variance computed on the effective rather than the nominal sample size. Each model is assessed against a sample stratified by its own map, so the stratum weights differ between rows. The design effect $\mathrm{deff} = n/n_{\mathrm{eff}}$ is the price of drawing validation pixels from 30 polygons instead of 1,829 independent locations; the hold-out column is repeated from Table 3 for contrast.

Model	Hold-out OA	Weighted OA	SE	95% CI	n	n_{eff}	deff
SVM (RBF)	0.9563	0.9701	0.0323	[0.9068, 1.0000]	1,829	25.8	70.8
LightGBM	0.9410	0.9574	0.0357	[0.8875, 1.0000]	1,829	24.4	74.9
MLP	0.9410	0.9572	0.0320	[0.8946, 1.0000]	1,829	24.3	75.3
XGBoost	0.9360	0.9484	0.0426	[0.8648, 1.0000]	1,829	24.6	74.2
Random Forest	0.9355	0.9583	0.0402	[0.8794, 1.0000]	1,829	24.2	75.5
CART	0.9267	0.9595	0.0328	[0.8952, 1.0000]	1,829	24.6	74.3

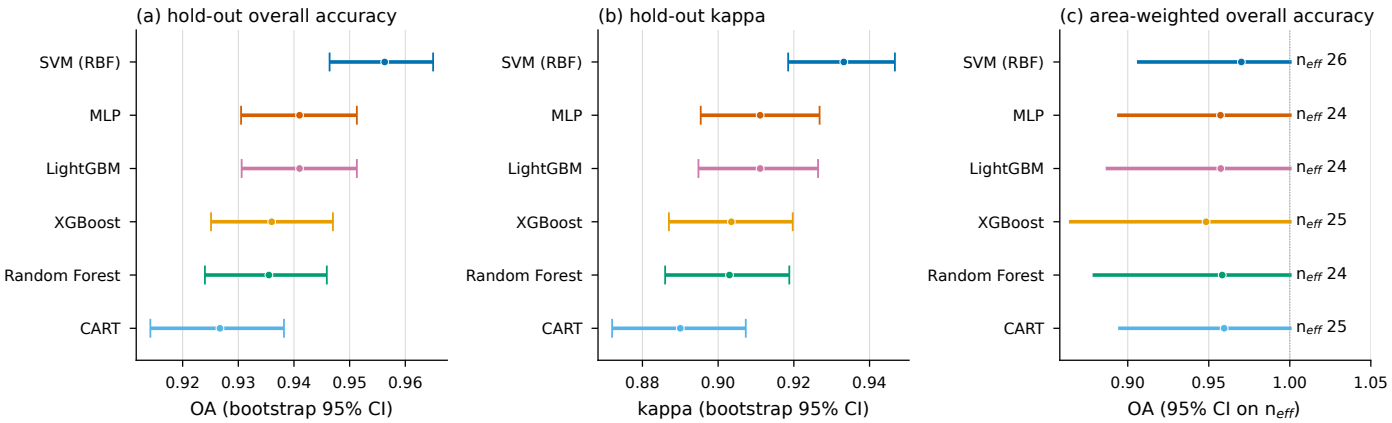


Figure 8. Model accuracy with uncertainty. (a) Hold-out overall accuracy and (b) hold-out κ, with 2,000-replicate bootstrap 95% intervals over the 1,829 validation pixels. (c) Area-weighted overall accuracy after Olofsson et al. [35], with the interval computed on the effective sample size that survives the polygon design effect; n_{eff} is annotated for each model.

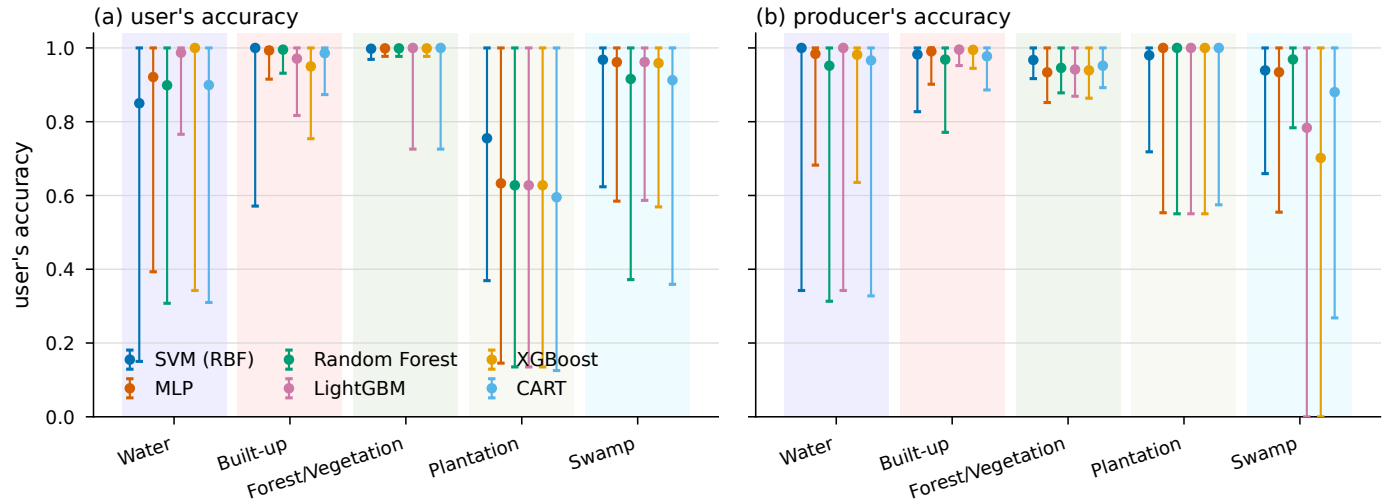


Figure 9. Per-class user's (a) and producer's (b) accuracy for each model, with 95% confidence intervals from the area-weighted estimator.

median single-model 95% margin of 15.15 km^2 (ratio 1.0): the choice of classifier moves the estimate as far as a typical single-model confidence interval allows. For swamp (0.7) and water (0.6) the between-model spread is somewhat smaller than the sampling margin, and for forest/vegetation (0.3) and plantation (0.3) it is much smaller.

In practice, an area quoted from a single classifier with its own confidence interval, which is the standard form of reporting in this literature, leaves out a between-model component that, for built-up land, is as large as the sampling margin itself and, for swamp

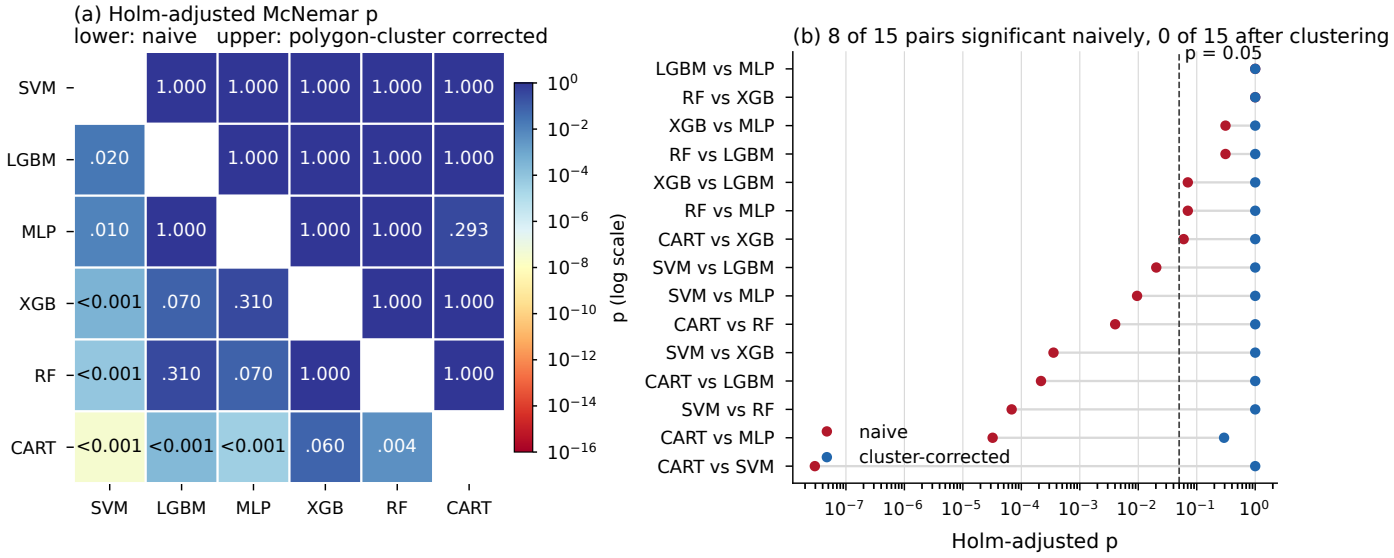


Figure 10. Pairwise significance. (a) Holm-adjusted McNemar p-values for all fifteen model pairs; the lower triangle is the standard pixel-level test, the upper triangle the cluster-corrected test of Durkalski et al. [21]. Colour is on a log scale. (b) The same fifteen comparisons as a paired dot plot, naive against cluster-corrected, with the 0.05 threshold marked.

Table 5. Mapped area of each class (km^2) in each of the six classified maps, with the between-model range and the ratio of the largest to the smallest estimate. Every map covers the same 270.44 km^2, so the columns differ only in how that area is apportioned.

Class	SVM (RBF)	LightGBM	MLP	XGBoost	Random Forest	CART	Range	Max/min
Water	14.53	8.93	9.93	7.98	7.42	7.66	7.12	1.96×
Built-up	71.68	95.87	83.70	96.03	83.55	102.06	30.38	1.42×
Forest/Vegetation	133.34	132.70	131.45	131.75	127.72	126.64	6.70	1.05×
Plantation	18.74	22.07	25.31	22.87	19.59	18.16	7.15	1.39×
Swamp	32.15	10.88	20.04	11.81	32.16	15.92	21.29	2.96×
Total	270.44	270.44	270.44	270.44	270.44	270.44	-	-

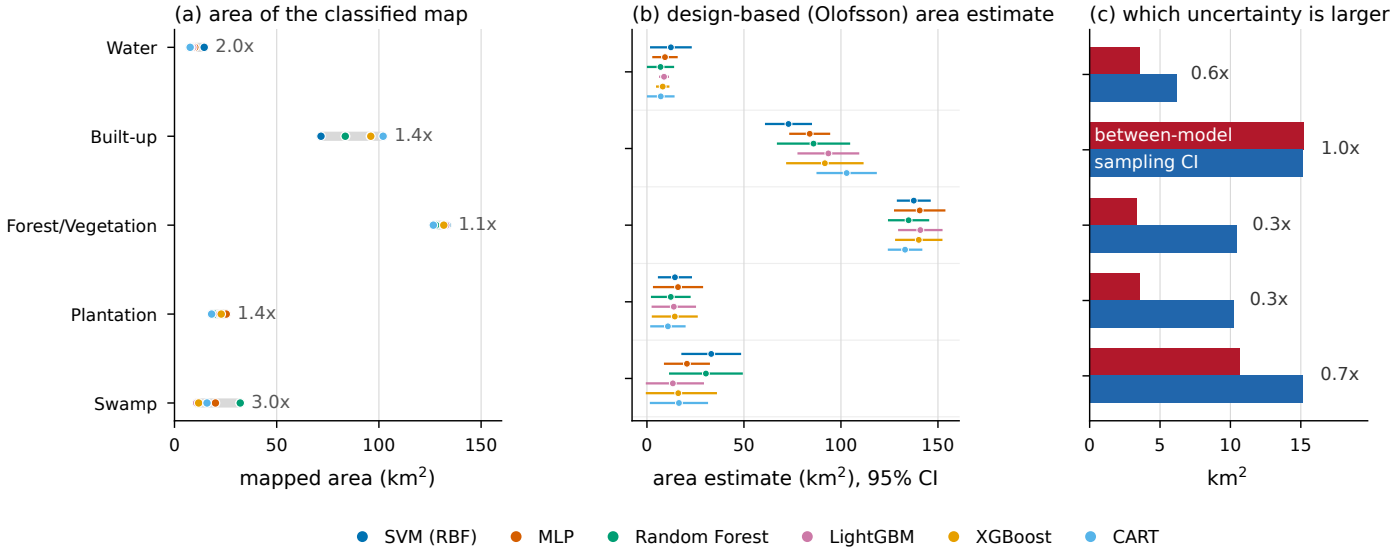


Figure 11. Between-model area uncertainty. (a) Mapped area of each class in each of the six maps, annotated with the ratio of largest to smallest. (b) Design-based area estimates with 95% intervals, computed per model after Olofsson et al. [35]. (c) For each class, half the between-model range (red) against the median single-model 95% half-width (blue); the annotated ratio is red divided by blue.

Table 6. Design-based area estimates for the most accurate model, SVM with an RBF kernel, after Olofsson et al. [35] on the effective sample size. The last three columns compare the two sources of uncertainty in an area figure: the sampling half-width is the median across the six models of that model's own 95% margin, and the model half-width is half the between-model range of Table 5. Where the ratio exceeds one, the choice of classifier moves the estimate further than the median single-model 95% margin.

Class	Mapped	Estimated	SE	95% CI	± sampling	± model	Ratio
Water	14.53	12.35	5.19	[2.18, 22.52]	6.20	3.56	0.6×
Built-up	71.68	72.96	5.88	[61.43, 84.49]	15.15	15.19	1.0×
Forest/Vegetation	133.34	137.55	4.16	[129.40, 145.70]	10.45	3.35	0.3×
Plantation	18.74	14.44	4.18	[6.24, 22.63]	10.26	3.58	0.3×
Swamp	32.15	33.14	7.56	[18.32, 47.96]	15.14	10.64	0.7×

and water, more than half of it.

4.5 Spatial and Spectral Structure of the Errors

The errors that remain are neither scattered nor model-specific, and three independent views converge on the same explanation.

First, the confusions recur across models. Table 7 lists every ordered class pair confused by at least one classifier. Forest/vegetation assigned to plantation affects all six models, with a median of 89 pixels, 9.18% of the class. Swamp assigned to water affects five models. Built-up assigned to swamp, and the reverse, affect six and five models respectively. Two pairs account for the great majority of all errors.

Second, within a class the errors concentrate in one or two polygons. Figure 12(b) shows that between 95.7% and 96.6% of every model's forest/vegetation errors fall inside a single validation polygon, number 77, out of the thirteen forest polygons in the validation set. For built-up, one polygon (number 16) accounts for 100% of the errors in XGBoost, LightGBM and the MLP and 88.9% in Random Forest. That six independently tuned classifiers from four model families fail on the same polygon points to a cause in the reference data rather than in the classifiers.

Third, those polygons sit in the wrong place in the feature space. Figure 12(c, d) projects each validation polygon onto the axis joining the centroids of its labelled class and the class it is confused with, in the z-scored four-band space. The offending polygons lie at or beyond the halfway point of that axis, that is, closer to the centroid of the other class than to their own. Polygon 77, a forest polygon, sits 2.59 times the class median distance from its own centroid. That is the signature of overlapping class definitions-a stand of managed trees is a legitimate member of either forest or plantation-rather than of mislabelled reference data.

The separability analysis is consistent with this interpretation. Table 8 and Figure 13 give the Jeffries-Matusita (JM) distance for all ten class pairs. Nine pairs exceed 1.98, comfortably separable. One does not: built-up against swamp, at 1.848. With only B2, B3, B4 and B8 available, deep street canyons in building shadow present the low reflectance across all four bands that also defines swamp. This is a property of the band set, not of any classifier, and it explains both why the boulevards of window B in Figure 7 are misassigned and why every one of the six models makes that same mistake.

4.6 Computational Cost

Because accuracy does not separate the six models, computational cost was also compared. Tuning time for the full 60-candidate randomised search spans more than two orders of magnitude (a factor of about 380), from 2.97 s for CART to 1,124.53 s for LightGBM, with the SVM at 6.44 s, XGBoost at 19.05 s, Random Forest at 54.65 s and the MLP at 98.93 s (Table 3). Inference over the full validation set orders the models differently: CART at 0.32 ms and the MLP at 1.08 ms are effectively free, while the SVM-the most accurate model-is the slowest at 106.23 ms, roughly 330 times CART, because kernel evaluation scales with the number of support vectors.

The implications of these costs for model choice are discussed in Section 5.2.

4.7 Toward an Independent Reference Sample

The accuracies above are estimated from a validation partition of the same purposive polygon set used for training. Splitting at polygon level removes the leakage between fitting and scoring, and weighting by mapped area removes the bias from disproportionate class representation, but neither converts a purposive sample into a probability sample. The estimates are conditional on the polygons having been placed

Table 7. Confusions recurring across models. For each ordered pair, the number of the six models affected, the median and range of the pixel count, and the median share of the true class assigned to the other class.

True class	Mapped as	Models	Median px	Range	% of class
Forest/Vegetation	Plantation	6	89	47-89	9.18
Swamp	Water	5	15	0-25	12.93
Built-up	Swamp	6	4	3-9	0.91
Swamp	Built-up	5	4	0-23	3.45
Water	Forest/Vegetation	3	0	0-1	0.33
Water	Built-up	2	0	0-1	0.00
Built-up	Water	1	0	0-2	0.00
Built-up	Plantation	2	0	0-13	0.00
Plantation	Forest/Vegetation	1	0	0-2	0.00

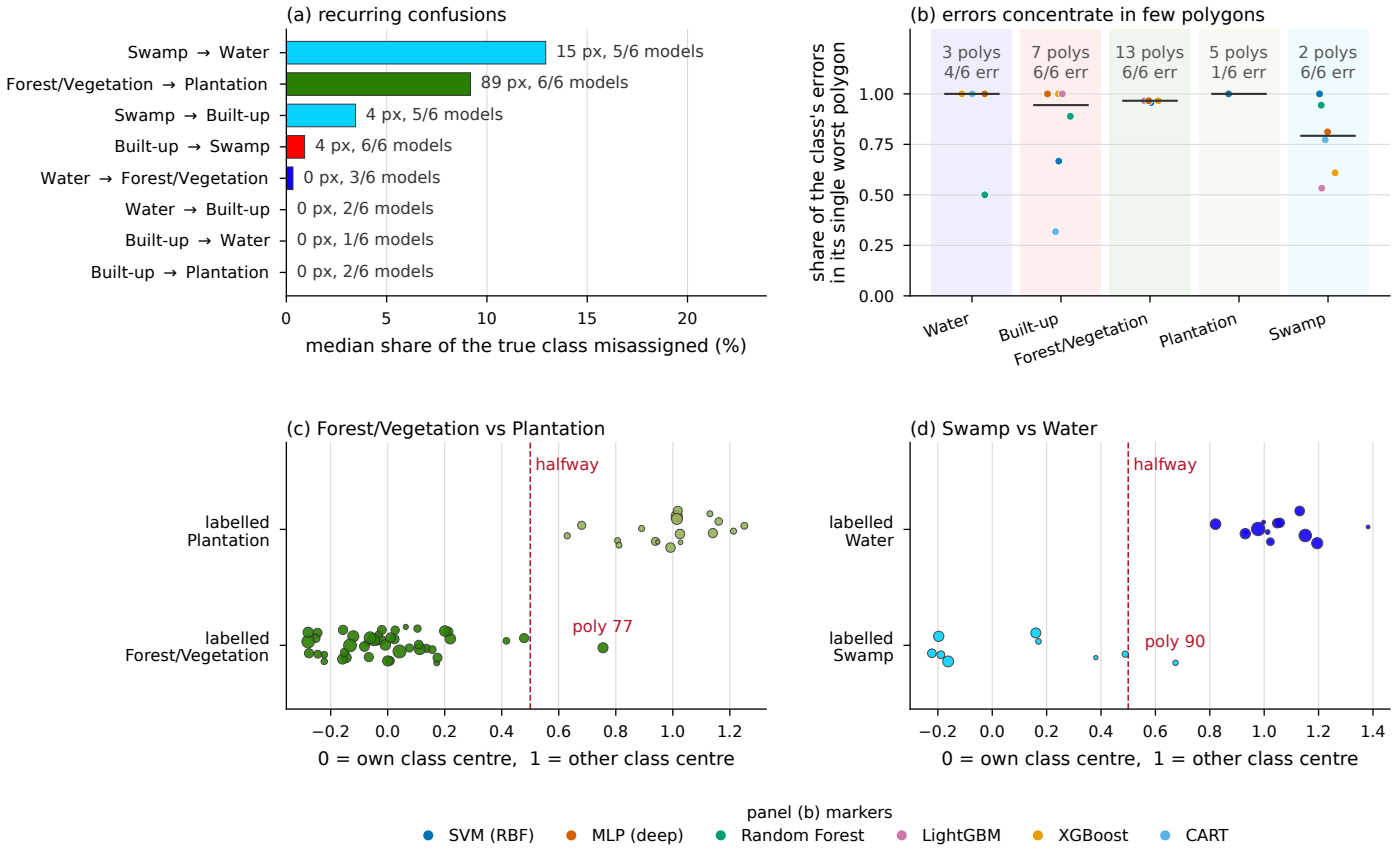


Figure 12. Where the misclassification is. (a) Recurring confusions: for each ordered class pair, the median across models of the share of the true class assigned to the other. (b) For each class and model, the share of that class's validation errors falling inside its single worst polygon. (c, d) Position of each validation polygon on the axis joining the two class centroids in the z-scored four-band feature space, where 0 is the centre of the labelled class and 1 the centre of the class it is confused with.

representatively, which cannot be verified from the sample itself. This is the reason the figures in Table 4 should be read as internally consistent comparisons across the six models rather than as unbiased estimates of map accuracy for Taipei.

A design for an independent reference sample is therefore given in the supplementary material (Table A4, Figure A1). Strata are the classes of the SVM map, weights W_i their area shares. Strictly proportional allocation of the 501-point base sample would give plantation 35 points, from which user's accuracy could be estimated no better than about ± 16 percentage points, so each rare class is raised to a floor of 75. Interpreting this sample against independent high-resolution imagery, and re-estimating accuracy

Table 8. Bhattacharyya and Jeffries-Matusita (JM) distances for the ten class pairs, computed on the training pixels in the four-band feature space. The measure is bounded above by 2 and pairs below roughly 1.9 are conventionally read as not reliably separable. Only Built-up against Swamp falls short.

Class A	Class B	Bhattacharyya	JM	Verdict
Built-up	Swamp	2.579	**1.8483**	marginal
Water	Built-up	4.978	1.9862	separable
Water	Swamp	5.077	1.9875	separable
Forest/Vegetation	Plantation	5.119	1.9880	separable
Plantation	Swamp	6.625	1.9973	separable
Built-up	Forest/Vegetation	10.592	1.9999	separable
Water	Forest/Vegetation	14.091	2.0000	separable
Water	Plantation	36.975	2.0000	separable
Built-up	Plantation	12.466	2.0000	separable
Forest/Vegetation	Swamp	20.705	2.0000	separable

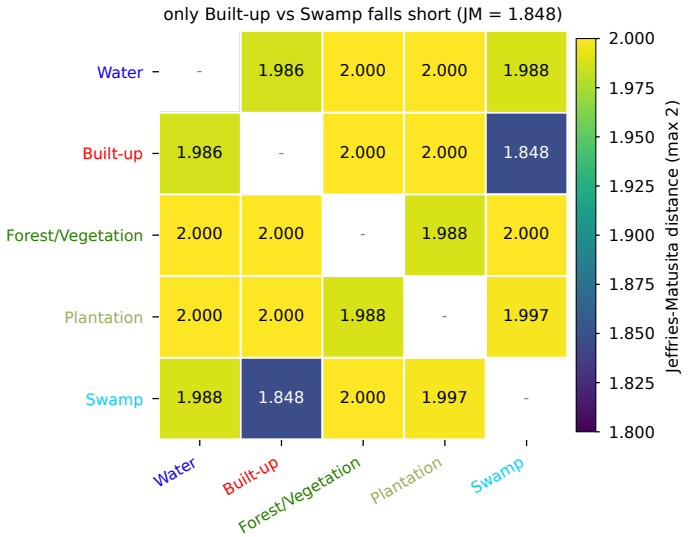


Figure 13. Pairwise class separability (JM distance) plotted against the conventional threshold of 1.9. Only built-up against swamp falls short, at 1.848.

and area from it, is the necessary next step and is beyond the scope of the present analysis.

5 Discussion

5.1 Why the Ensembles Do Not Outperform a Single Tree

The expected result of a six-way comparison of this kind is that the boosted ensembles and the SVM outperform a single pruned tree by a clear margin. They do not. CART reaches 0.9267 against 0.9563 for the SVM, and LightGBM, given about 380 times CART's tuning budget in wall-clock terms, reaches 0.9410.

The competitiveness of a single pruned tree relative to more complex models is not without precedent: Tayyebi and Pijanowski [38] found that CART offered a favourable trade-off between goodness of fit and interpretability compared with ANN and MARS in a land use change modelling context. The separability analysis explains why the same pattern emerges here. Nine of the ten class pairs are separable at JM > 1.98 in the four-band space; for those pairs a single pruned tree with five levels finds essentially the same decision surface as a kernel machine, because the classes are close to linearly separable and there is little structure left for a more expressive model to exploit. The tenth pair, built-up against swamp at JM 1.848, is not separable by *any* decision rule operating on these four bands, because the two classes occupy overlapping regions of the space. Model capacity cannot recover information that the feature vector does not contain. The result is a ceiling that all six models approach and none exceeds.

This reframes a common finding in the GEE benchmarking literature. Studies that report Random Forest superiority [3, 7, 11, 13] typically do so on margins of a few accuracy points, computed on pixel-level splits of polygon-derived samples, without a significance test or with an uncorrected one. Our results do not contradict those point estimates; they suggest that margins of that size, on samples of that structure, may not be distinguishable from zero. The appropriate response is not to abandon comparative studies but to report the effective sample size alongside the nominal one, and to test differences with a method that respects the clustering.

5.2 Cost as the Deciding Criterion

If the six models cannot be separated on accuracy, the choice must rest on other grounds, and computational cost is the most tractable. The ordering by cost is not the ordering by accuracy, and it is not even consistent between the two cost measures. CART is cheapest on both, at 2.97 s to tune and 0.32 ms to apply. The SVM is cheap to tune (6.44 s) but by far the most expensive to apply (106.23 ms), because RBF kernel evaluation scales with the number of support vectors, which grows with training set size. LightGBM is the most expensive to tune, at 1,124.53 s, about eleven times the MLP (98.93 s), and the second most expensive to apply (40.24 ms).

For a single city map any of these is acceptable. For an operational service reclassifying a large archive, or for a wall-to-wall product updated seasonally, inference cost dominates; a 330-fold difference between two models that cannot be statistically separated then favours CART or the MLP. For a one-off study where a map is produced once and inspected many times, tuning cost is what the analyst waits for, and LightGBM is the poor choice. The general point is that once accuracy ceases to discriminate, the selection criterion should be stated explicitly and matched to the intended deployment, rather than inherited from a ranking in a previous paper.

5.3 Errors as a Property of the Reference Data

The concentration of errors in individual polygons is the most operationally useful finding here. That 96% of every model's forest errors fall in a single polygon out of thirteen, and that this polygon sits 2.59 times the class median distance from its own class centroid, says that the marginal return on a better classifier is close to zero while the marginal return on revisiting that one polygon is large.

It also bears on how such polygons should be treated. The temptation is to label them errors in the reference data and remove them. The feature-space projection argues against that: polygon 77 lies past the halfway point on the forest-plantation centroid axis, which is what a genuine intermediate-a stand of managed trees on a hillslope-looks like. Removing it would raise every model's accuracy without improving any map. The more defensible response is to refine the class definition, or to accept that the boundary between forest and plantation is not a boundary the four-band feature space can resolve, and report the two classes merged where that is acceptable to the application.

5.4 Implications for Planning Use

The area results carry the practical consequence. Built-up land in Taipei is mapped at anywhere between 71.68 and 102.06 km^2 depending on which of six statistically indistinguishable classifiers is used, a range of 30.38 km^2 that exceeds the entire mapped extent of water and, in most maps, of plantation and swamp. A planning application that reads an impervious-surface figure off one such map and applies it to runoff modelling, heat exposure or infrastructure demand is working with a number whose uncertainty is larger than the map's own accuracy assessment reports, by a factor of about 1.4 if the between-model and sampling components are combined in quadrature.

The remedy is not to search for the correct classifier, since the data do not identify one. It is to report an ensemble range. Where a single figure is required, the design-based estimate of Table 6 with an interval widened to cover the between-model spread is the more defensible form. We therefore report built-up land for Taipei in 2020 as a range across models rather than as a point estimate from the most accurate one.

The consequence is sharpest for hazard applications. Taipei's flood exposure is governed by impervious cover and by the condition of the riparian margins along the Tamsui and Keelung, and both quantities are classifier-dependent here (Table 5). A runoff or inundation model calibrated on one classifier's impervious fraction inherits that spread as an unstated boundary condition, and a typhoon-season exposure figure derived from one map carries an uncertainty that the map's own accuracy assessment does not disclose. Where a classification is to drive a hazard model rather than a descriptive inventory, the case for propagating the between-model range instead of a single surface is correspondingly stronger.

5.5 Transferability to Other Landscapes

A single-city result invites the question of what it predicts elsewhere, and the mechanism identified above makes that question answerable rather than speculative. The null result here is not a property of Taipei. It is a property of a feature space in which nine of ten class pairs are separable at JM > 1.98, and it follows from a simple condition: where the classes are close to separable in the bands available, every classifier with enough capacity to find the boundary converges on much the same one, and the differences between them fall below what a clustered sample can resolve. Model choice should then be immaterial.

Where classes overlap substantially in the feature space, capacity has something to exploit, and the models should separate.

This prediction can be tested in both directions. In a landscape with a coarse class scheme, a wide band set, or classes as spectrally distinct as open water against dense urban fabric, we would expect the same null: the boosted ensembles will not beat a pruned tree by a margin that survives correction. In a landscape with a fine class scheme on a minimal band set-smallholder crop types, informal against low-rise formal settlement, degraded against intact forest-the pairwise separabilities will sit well below 1.98 and a capacity-driven ranking should re-emerge. The diagnostic is available before any classifier is fitted: compute the pairwise separability of the intended class set in the intended feature space, and the answer to whether the algorithm matters is largely determined. We would encourage that calculation to be reported in comparative studies, since it costs almost nothing and tells the reader in advance whether an accuracy ranking is informative or noise.

The sampling result transfers more widely than the accuracy result. Any study that digitises training polygons and then splits them at pixel level carries the design effect of Eq. 5, whatever the city, sensor or class scheme; here it reduced 1,829 validation pixels to between 24 and 26 independent observations. That is a property of the sample design rather than of Taipei, and the correction applied here should be applied wherever the design is the same. A replication on a second, structurally different city-one with a lower separability floor-would test the accuracy claim directly, and is the extension we regard as most informative.

5.6 Limitations

Several limitations bound these conclusions.

The reference sample is purposive, not probabilistic. The 110 polygons were hand-digitised to cover the classes, not drawn at random from the study area, so the accuracy figures are conditional on those polygons being spectrally representative of their classes across the city. Splitting at polygon level removes leakage between fitting and scoring, and the Olofsson weighting removes bias from disproportionate class representation, but neither addresses this. The independent probability sample designed in Table A4 is the correct instrument and its interpretation is outstanding.

The effective sample size is small in absolute terms. Some 24 to 26 independent observations are enough to establish that the six models are not distinguishable, but not enough to establish that they are equivalent. Absence of a significant difference is not evidence of no difference; a larger, better-distributed polygon set could in principle separate them. What the present data show is that the existing evidence base does not.

The feature set is deliberately minimal, and it is top-of-atmosphere. Four Level-1C bands were used to isolate the effect of the classifier. Repeating the comparison on Level-2A surface reflectance would very likely shift the absolute accuracies; we would not expect it to change the ranking, since the atmosphere acts on all six models alike, but that is an assumption rather than a result. Adding the red-edge and shortwave-infrared bands (B5-B7, B11, B12), spectral indices such as the normalised difference vegetation index (NDVI), normalised difference built-up index (NDBI) and modified normalised difference water index (MNDWI), texture measures or Sentinel-1 backscatter [2, 39, 40] would very likely raise the built-up-swamp separability above the 1.9 threshold and remove the boulevard artefact. It might also re-separate the models, since a richer feature space gives capacity something to exploit. That is a testable prediction and a natural extension, but it is a different experiment from the one reported here.

The composite is annual, and that choice has consequences of its own. A full-year median was used because Taipei's summer cloud regime leaves too few clear scenes for a stable seasonal composite, but a median taken over twelve months averages the phenological cycle away. Plantation and forest/vegetation are precisely the two classes whose separation depends on that cycle, since managed vegetation follows a cropping calendar the hill forest does not, and the annual median discards the signal that would distinguish them. The forest-to-plantation confusion that affects all six models (Table 7) is therefore partly an artefact of the compositing window rather than of the classifier or the band set alone. A seasonal composite stack, or harmonic coefficients fitted to the annual reflectance series, would test that directly, and of the extensions listed here it is the one most likely to change the result.

The comparison covers a single year. The transferability argument of Section 5.5 predicts what should happen in a different landscape, but says nothing about time series, where inter-annual

consistency rather than single-date accuracy is the criterion and a model that is marginally less accurate but more stable between years may well be preferable. Nothing in these results speaks to that trade-off.

Finally, the classification is per-pixel. Object-based approaches, which classify segments rather than pixels, change the unit of analysis and with it the clustering structure of the validation sample; the corrections applied here would need to be reformulated rather than simply carried over.

In order of priority, the next steps are: (1) interpreting the 605-point probability sample of Table A4, which removes the purposive-sampling caveat; (2) refitting the models on a seasonal or harmonic composite stack; (3) widening the feature space as described above; (4) replicating on a second city with a lower separability floor; and (5) extending the design to a multi-year series and to object-based units. Only the first is required to support the present conclusions; the others test them.

6 Conclusions

Six machine learning classifiers-CART, Random Forest, XGBoost, LightGBM, a multilayer perceptron and an RBF support vector machine-were tuned and compared for LULC mapping of Taipei City on a four-band Sentinel-2 composite, under a partitioning scheme that keeps every hand-digitised polygon whole through both the hold-out split and the tuning folds.

Hold-out overall accuracy spans 2.96 percentage points, from 0.9267 to 0.9563, and the bootstrap interval of the SVM does not overlap those of Random Forest and CART. This apparent separation is not supported once the sampling design is taken into account. The 1,829 validation pixels come from 30 polygons and carry an effective sample size of only 24 to 26 independent observations. Eight of the fifteen pairwise McNemar comparisons are significant when the pixels are treated as independent; none is significant once the clustering is corrected for.

The six maps are nonetheless not interchangeable. Mapped built-up land ranges from 71.68 to 102.06 km^2 and swamp by a factor of 2.96, and for built-up land half the between-model range equals the median 95% half-width of a single model's area estimate, while for swamp and water it reaches 70% and 60% of it. Accuracy and area therefore give different answers to the question of whether the classifier matters: on accuracy it does not, on area it does.

The errors that remain are structural rather than algorithmic. Two class pairs account for the great majority of them; within a class they concentrate in one or two polygons; and those polygons sit past the halfway point of the axis joining the two class centroids. Built-up against swamp is the one class pair whose Jeffries-Matusita distance falls short of separability, at 1.848, which is a property of the four-band feature set and not of any decision rule.

Three recommendations follow for comparative LULC studies. Report the effective sample size alongside the nominal one, since the two differ here by a factor of 71 to 76. Test differences between models with a method that accounts for clustering in the reference data, since the naive test rejected at 82.5% under a simulated clustered null at $\alpha = 0.05$. And where areas are the product, report the between-model range alongside a single model's design-based interval, since for built-up land the two are of the same size. Where a classifier must nonetheless be chosen, computational cost is the defensible criterion: CART is roughly 330 times faster at inference than the most accurate model, from which it cannot be statistically distinguished.

Data Availability Statement

Sentinel-2 imagery is freely available through the Google Earth Engine data catalog as `COPERNICUS/S2_HARMONIZED`. The Earth Engine scripts used for compositing, sampling and export, the local analysis pipeline including the polygon-grouped splitting, the tuning, the clustered McNemar implementation and the area-weighted estimator, the 110 training polygons, the six classified GeoTIFFs, and the scripts that generate every table and figure in this paper are openly available at https://github.com/chunyaliu1115/Taipei_LULC. Code is released under the MIT licence and the data under CC BY 4.0. The band values at all 8,493 labelled pixels are included, so every accuracy, significance and separability result reported here can be reproduced without an Earth Engine account. All tables are regenerated from the analysis outputs by a single script.

Funding

The author acknowledges Google LLC for providing the Google Earth Engine platform and the European Space Agency for the Copernicus Sentinel-2 data (contains modified Copernicus Sentinel data, 2020), and thanks the anonymous reviewers for their

comments on validation design. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The author declares no conflicts of interest. The author holds no financial or proprietary interest in any of the software, platforms or data sources used in this study, and received no funding, compensation or other consideration from any commercial or governmental entity in connection with this work.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study used only satellite imagery and map-derived reference data and involved no human participants or animals.

References

[1] Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., ... & Snyder, P. K. (2005). Global consequences of land use. *science, 309*(5734), 570-574. [CrossRef]

[2] Mishra, V. N., Prasad, R., Rai, P. K., Vishwakarma, A. K., & Arora, A. (2019). Performance evaluation of textural features in improving land use/land cover classification accuracy of heterogeneous landscape using multi-sensor remote sensing data. *Earth Science Informatics, 12*(1), 71-86. [CrossRef]

[3] Zhao, Z., Islam, F., Waseem, L. A., Tariq, A., Nawaz, M., Islam, I. U., ... & Hatamleh, W. A. (2024). Comparison of three machine learning algorithms using google earth engine for land use land cover classification. *Rangeland ecology & management, 92*, 129-137. [CrossRef]

[4] Ouchra, H., Belangour, A., Erraissi, A., & Banane, M. (2023, September). Assessing machine learning algorithms for land use and land cover classification in Morocco using google earth engine. In *International Conference on Image Analysis and Processing* (pp. 395-405). Cham: Springer Nature Switzerland. [CrossRef]

[5] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment, 202*, 18-27. [CrossRef]

[6] Gumma, M. K., Thenkabail, P. S., Teluguntla, P. G., Oliphant, A., Xiong, J., Giri, C., ... & Whitbread, A. M. (2020). Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud. *GIScience & Remote Sensing, 57*(3), 302-322. [CrossRef]

[7] Pande, C. B. (2022). Land use/land cover and change detection mapping in Rahuri watershed area (MS), India using the google earth engine and machine learning approach. *Geocarto International, 37*(26), 13860-13880. [CrossRef]

[8] Feizizadeh, B., Omarzadeh, D., Kazemi Garajeh, M., Lakes, T., & Blaschke, T. (2023). Machine learning data-driven approaches for land use/cover mapping and trend analysis using Google Earth Engine. *Journal of Environmental Planning and Management, 66*(3), 665-697. [CrossRef]

[9] Shao, Y., & Lunetta, R. S. (2012). Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS Journal of Photogrammetry and Remote Sensing, 70*, 78-87. [CrossRef]

[10] Qian, Y., Zhou, W., Yan, J., Li, W., & Han, L. (2014). Comparing machine learning classifiers for object-based land cover classification using very high resolution imagery. *Remote sensing, 7*(1), 153-168. [CrossRef]

[11] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing, 67*, 93-104. [CrossRef]

[12] Ghosh, A., Sharma, R., & Joshi, P. K. (2014). Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion. *Applied Geography, 48*, 31-41. [CrossRef]

[13] Arpitha, M., Ahmed, S.A., & Harishnaika, N. (2023). Land use and land cover classification using machine learning algorithms in Google Earth Engine. *Earth Science Informatics, 16*(4), 3057-3073. [CrossRef]

[14] Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote sensing, 12*(7), 1135. [CrossRef]

[15] Camargo, F. F., Sano, E. E., Almeida, C. M., Mura, J. C., & Almeida, T. (2019). A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images. *Remote Sensing, 11*(13), 1600. [CrossRef]

[16] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic

structure. *Ecography, 40*(8), 913-929. [CrossRef]

[17] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., ... & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications, 11*(1), 4540. [CrossRef]

[18] Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications, 13*(1), 2208. [CrossRef]

[19] Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York. [CrossRef]

[20] Foody, G. M. (2004). Thematic map comparison. *Photogrammetric Engineering & Remote Sensing, 70*(5), 627-633. [CrossRef]

[21] Durkalski, V. L., Palesch, Y. Y., Lipsitz, S. R., & Rust, P. F. (2003). Analysis of clustered matched-pair data. *Statistics in medicine, 22*(15), 2417-2428. [CrossRef]

[22] Jen-Chieh, C., & Yu-i, L. (2013, August). Using Formosat-2 Satellite Images to Manage Hillside Land Cover Change in Taipei City. In *2013 the International Conference on Remote Sensing, Environment and Transportation Engineering (RSETE 2013)* (pp. 849-852). Atlantis Press. [CrossRef]

[23] Liu, C.-Y. & Arora, A. (2024). Decadal analysis and simulation of land use and land cover changes in Taiwan using machine learning and Markov chain models. *Environment, Development and Sustainability, 28*, 22577-22613. [CrossRef]

[24] Chen, Y. Y., Huang, W., Wang, W. H., Juang, J. Y., Hong, J. S., Kato, T., & Luyssaert, S. (2019). Reconstructing Taiwan's land cover changes between 1904 and 2015 from historical maps and satellite images. *Scientific reports, 9*(1), 3643. [CrossRef]

[25] Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA. [CrossRef]

[26] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. [CrossRef]

[27] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing, 26*(1), 217-222. [CrossRef]

[28] 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*, 785-794. [CrossRef]

[29] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems, 30*, 3146-3154.

[30] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. [CrossRef]

[31] Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS journal of photogrammetry and remote sensing, 66*(3), 247-259. [CrossRef]

[32] Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. Technical Report, Department of Computer Science, National Taiwan University, Taipei. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[33] Martins, S., Bernardo, N., Ogashawara, I., & Alcantara, E. (2016). Support vector machine algorithm optimal parameterization for change detection mapping in Funil Hydroelectric Reservoir (Rio de Janeiro State, Brazil). *Modeling Earth Systems and Environment, 2*(3), 138. [CrossRef]

[34] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46. [CrossRef]

[35] Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote sensing of Environment, 148*, 42-57. [CrossRef]

[36] Stehman, S. V. (2014). Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *International journal of remote sensing, 35*(13), 4923-4939. [CrossRef]

[37] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. https://www.ime.usp.br/~abe/lista/pdf4R8xPVzCnX.pdf

[38] Tayyebi, A., & Pijanowski, B. C. (2014). Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools. *International journal of applied earth observation and geoinformation, 28*, 102-116. [CrossRef]

[39] Nigar, A., Li, Y., Jat Baloch, M. Y., Alrefaei, A. F., & Almutairi, M. H. (2024). Comparison of machine and deep learning algorithms using Google Earth Engine and Python for land classifications. *Frontiers in Environmental Science, 12*, 1378443. [CrossRef]

[40] Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International journal of remote sensing, 24*(3), 583-594. [CrossRef]

Appendix

The four tables and one figure below support the main text without being needed to follow its argument. Tables A1 and A2 are the full numeric backing for results the main text states in summary or plots in Figure 9; Table A3 gives all fifteen pairwise tests behind Figure 10; and Figure A1 with Table A4 specify the probability sample proposed as the next step, which is a design rather than a result of this study. They are numbered in the order they are first cited above.

Table A1. Hyperparameters selected by five-fold polygon-grouped randomised search, 60 candidates per model.

Model	Selected values
SVM (RBF)	C = 20.6914, gamma = 0.000615848
LightGBM	colsample_bytree = 1, learning_rate = 0.05, min_child_samples = 10, n_estimators = 800, num_leaves = 15, reg_lambda = 1, subsample = 0.8
MLP	alpha = 0.0359381, hidden_layer_sizes = [64], learning_rate_init = 0.001
XGBoost	colsample_bytree = 1, learning_rate = 0.05, max_depth = 9, min_child_weight = 1, n_estimators = 200, reg_lambda = 5, subsample = 0.6
Random Forest	class_weight = balanced_subsample, max_depth = 30, max_features = sqrt, min_samples_leaf = 10, n_estimators = 50
CART	ccp_alpha = 0.0001, criterion = gini, max_depth = 5, min_samples_leaf = 1, min_samples_split = 5

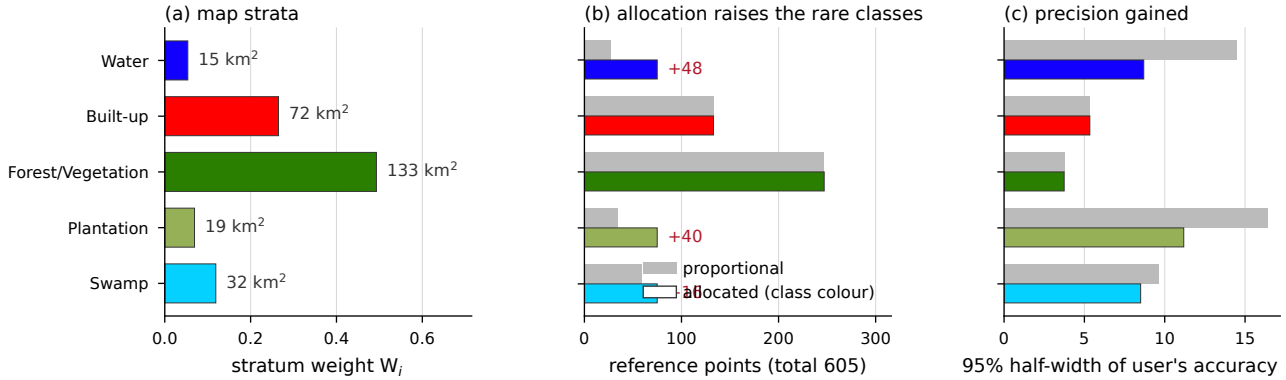


Figure A1. Design of the independent reference sample. (a) Stratum weights W_i. (b) Proportional allocation against the allocation used, which raises each rare class to a floor of 75 points. (c) The resulting 95% half-width of the user's accuracy estimate at the prior accuracy assumed for each class.

Table A2. Numerical values of the area-weighted user's accuracy (UA) and producer's accuracy (PA) per class and model, with 95% confidence intervals on the effective sample size.

	UA	95% CI	PA	95% CI
Water				
SVM (RBF)	0.850	[0.150, 1.000]	1.000	[0.342, 1.000]
LightGBM	0.987	[0.766, 1.000]	1.000	[0.342, 1.000]
MLP	0.921	[0.393, 1.000]	0.984	[0.682, 1.000]
XGBoost	1.000	[0.342, 1.000]	0.982	[0.635, 1.000]
Random Forest	0.899	[0.308, 1.000]	0.952	[0.313, 1.000]
CART	0.899	[0.310, 1.000]	0.966	[0.328, 1.000]
Built-up				
SVM (RBF)	1.000	[0.571, 1.000]	0.982	[0.827, 1.000]
LightGBM	0.971	[0.817, 1.000]	0.996	[0.952, 1.000]
MLP	0.993	[0.915, 1.000]	0.991	[0.902, 1.000]
XGBoost	0.950	[0.754, 1.000]	0.995	[0.945, 1.000]
Random Forest	0.995	[0.931, 1.000]	0.968	[0.771, 1.000]
CART	0.986	[0.873, 1.000]	0.977	[0.886, 1.000]
Forest/Vegetation				
SVM (RBF)	0.998	[0.969, 1.000]	0.967	[0.917, 1.000]
LightGBM	1.000	[0.726, 1.000]	0.942	[0.869, 1.000]
MLP	0.999	[0.977, 1.000]	0.934	[0.852, 1.000]
XGBoost	0.999	[0.977, 1.000]	0.939	[0.864, 1.000]
Random Forest	0.999	[0.977, 1.000]	0.946	[0.878, 1.000]
CART	1.000	[0.726, 1.000]	0.952	[0.892, 1.000]
Plantation				
SVM (RBF)	0.755	[0.369, 1.000]	0.980	[0.718, 1.000]
LightGBM	0.628	[0.135, 1.000]	1.000	[0.550, 1.000]
MLP	0.633	[0.145, 1.000]	1.000	[0.553, 1.000]
XGBoost	0.628	[0.135, 1.000]	1.000	[0.550, 1.000]
Random Forest	0.628	[0.135, 1.000]	1.000	[0.550, 1.000]
CART	0.595	[0.125, 1.000]	1.000	[0.575, 1.000]
Swamp				
SVM (RBF)	0.968	[0.624, 1.000]	0.939	[0.659, 1.000]
LightGBM	0.962	[0.587, 1.000]	0.783	[0.000, 1.000]
MLP	0.962	[0.585, 1.000]	0.934	[0.555, 1.000]
XGBoost	0.959	[0.569, 1.000]	0.702	[0.000, 1.000]
Random Forest	0.916	[0.372, 1.000]	0.969	[0.784, 1.000]
CART	0.913	[0.359, 1.000]	0.880	[0.268, 1.000]

Table A3. Holm-adjusted p-values of the naive and cluster-corrected McNemar tests for all fifteen model pairs, adjusted within each column. The naive column treats the validation pixels as independent; the clustered column applies the correction of Durkalski et al. [21], which aggregates discordant outcomes within polygon before squaring. Eight of the fifteen differences are significant at $\alpha = 0.05$ under the naive test and none is significant once the clustering is accounted for.

Model A	Model B	ΔOA	b	c	Disc. polys	p_{naive}	$p_{\text{clustered}}$	Sig. (clust.)
CART	SVM (RBF)	-0.0295	66	12	10	2.94e-08	1.000	no
CART	Random Forest	-0.0087	18	2	6	4.02e-03	1.000	no
CART	XGBoost	-0.0093	27	10	8	5.97e-02	1.000	no
CART	LightGBM	-0.0142	30	4	7	2.17e-04	1.000	no
CART	MLP	-0.0142	27	1	9	3.23e-05	0.293	no
SVM (RBF)	Random Forest	+0.0208	14	52	10	6.83e-05	1.000	no
SVM (RBF)	XGBoost	+0.0202	19	56	9	3.55e-04	1.000	no
SVM (RBF)	LightGBM	+0.0153	26	54	8	2.03e-02	1.000	no
SVM (RBF)	MLP	+0.0153	20	48	9	9.54e-03	1.000	no
Random Forest	XGBoost	-0.0005	14	13	5	1.00e+00	1.000	no
Random Forest	LightGBM	-0.0055	18	8	6	3.10e-01	1.000	no
Random Forest	MLP	-0.0055	12	2	6	7.03e-02	1.000	no
XGBoost	LightGBM	-0.0049	10	1	3	7.03e-02	1.000	no
XGBoost	MLP	-0.0049	15	6	3	3.10e-01	1.000	no
LightGBM	MLP	+0.0000	8	8	4	1.00e+00	1.000	no

Table A4. Planned allocation of the independent probability reference sample, stratified by the classes of the SVM map. Strictly proportional allocation would leave the rare classes too thin to estimate a user's accuracy from, so each is raised to a floor of 75 points.

Stratum	Area (km^2)	W_i	Prior UA	Proportional	Allocated
Water	14.53	0.0537	0.820	27	75
Built-up	71.68	0.2650	0.889	133	133
Forest/Vegetation	133.34	0.4931	0.900	247	247
Plantation	18.74	0.0693	0.576	35	75
Swamp	32.15	0.1189	0.830	59	75
Total	270.44	1.0000	-	501	605

Chun-ya Liu is affiliated with Géographie-cités, University of Paris 1 Panthéon-Sorbonne, 93322, Aubervilliers, France. Her research focuses on land use and land cover change, urban environmental assessment, and the spatial modeling of landscape dynamics, with applications in East Asian study areas. (Email: chunyaliu@hotmail.com)