A Scalable Open-Source Workflow for Riverbed Substrate Classification Using UAV Imagery

Tulio Soto Parra , David Farò , Guido Zolezzi *

Posted Date: 21 May 2026

doi: 10.20944/preprints202605.1410.v1

Keywords: unmanned aerial systems; sediment mapping; spatial autocorrelation; ecohydraulics; hyperspatial resolution

 doi:10.20944/preprints202605.1410.v1

Article

A Scalable Open-Source Workflow for Riverbed Substrate Classification Using UAV Imagery

Tulio Soto Parra [1,2], David Farò [2,3] and Guido Zolezzi [3,*]

[1] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana
[2] Department of Civil, Environmental, and Mechanical Engineering, University of Trento, Via Mesiano 77, 38123 Trento, Italy
[3] Leibniz Institute of Freshwater Ecology and Inland Fisheries, IGB Berlin, Müggelseedamm 310, 12587 Berlin, Germany
* Correspondence: guido.zolezzi@unitn.it

Abstract

Accurate characterization of riverbed substrate from remote sensing imagery is essential for applications in fluvial geomorphology, habitat modeling, and river management. While recent advances in computer vision, particularly deep learning, have improved sediment mapping capabilities, their reliance on large annotated datasets and computational resources limits their broader applicability. This study presents a scalable workflow for categorical substrate classification using ultra-high-resolution UAV-derived RGB orthoimagery in clear-water river environments. The approach integrates spectral information with statistical and structural texture descriptors derived from Gray-Level Co-occurrence Matrices (GLCM) and Local Binary Patterns (LBP), combined within a Random Forest classification framework. The methodology consists of two main steps: (i) manual annotation of homogeneous substrate patches within a standard GIS environment and (ii) automated feature extraction, model optimization, and full-domain classification. Model performance is evaluated using spatially aware cross-validation and design-based probability sampling to account for spatial autocorrelation and provide unbiased accuracy estimates. The method was applied in four geomorphologically distinct alpine river reaches, achieving design-based overall accuracy ranging from 69.77% to 95.21%. These results demonstrate that RGB-based approaches can achieve reliable reach-scale categorical substrate classification when combined with appropriate feature representation and rigorous validation strategies. However, limitations remain for visually similar or transitional substrate classes, particularly fine sediments such as sand and clay, which are difficult to distinguish consistently even during manual annotation. The workflow is implemented using open-source tools and is applicable to clear-water conditions where the riverbed remains optically visible.

Keywords: unmanned aerial systems; sediment mapping; spatial autocorrelation; ecohydraulics; hyperspatial resolution

1. Introduction

The characterization of riverbed sediment composition is a fundamental requirement for fluvial geomorphology, ecohydraulics, and effective river management. Sediment grain size directly influences hydraulic roughness, sediment transport thresholds, and the formation and persistence of aquatic habitats, thus regulating key geomorphic and ecological processes Graham et al. (2005); Woodget and Austrums (2017). Historically, these parameters have been quantified using field-based techniques such as the Wolman pebble count Wolman (1954), which, despite their usefulness, are inherently limited by sparse sampling density and the inability to capture the high spatial heterogeneity of dynamic gravel-bed rivers Carbonneau et al. (2004b); Rogers et al. (2025).

The emergence of Unmanned Aerial Vehicles (UAVs) and Structure-from-Motion (SfM) photogrammetry has enabled a paradigm shift toward continuous, spatially explicit characterization of riverbed sediment Langhammer and Vacková (2018); Woodget and Austrums (2017). Early remote

 doi:10.20944/preprints202605.1410.v1

sensing approaches primarily focused on estimating sediment composition as a continuous variable (e.g., D_{50}), using either object-based image analysis or statistical relationships between image texture and grain size Buscombe (2013); Carbonneau et al. (2004b); Verdú et al. (2005). More recently, deep learning–based approaches have been introduced to automate grain-size estimation from imagery, demonstrating improved robustness and transferability Buscombe (2020); Mair et al. (2024). While these methods provide detailed descriptions of grain size distributions, they are often constrained by high spatial resolution requirements, sensitivity to environmental conditions, the need for site-specific calibration, and, in some cases, by the requirement of large manually labelled datasets and of substantial computational resources Carbonneau et al. (2005); Lang et al. (2020); Purinton and Bookhagen (2019). As a result, their applicability at the reach scale remains limited Farò et al. (2025).

Many practical applications, particularly in river habitat modeling and geomorphic assessment, nevertheless rely on categorical representations of sediment size, where substrates are grouped into discrete size classes such as sand, gravel, cobbles, and boulders Farò et al. (2025); Knighton (2014); Vezza et al. (2014); Woodget et al. (2016). Although categorical classification does not provide the same level of detail as continuous grain size estimation, it offers several practical advantages: it reduces data and resolution requirements, is usually less sensitive to environmental variability, and enables direct mapping of spatially coherent substrate patches relevant to ecological and hydraulic processes Carbonneau et al. (2004b); Farò et al. (2025); Woodget et al. (2016). These characteristics make categorical approaches particularly suitable for reach-scale and operational applications such as habitat assessment, geomorphic characterization, sediment monitoring, and river restoration planning.

Despite this potential, developing scalable and reliable workflows for categorical substrate classification has not been widely developed so far. RGB-based approaches, which rely on spectral and texture features derived from optical imagery, offer a practical and accessible solution Arif et al. (2017); Giroux et al. (2022). However, their performance is still affected by environmental variability, including illumination conditions, water depth, and sediment heterogeneity, and typically rely on proprietary software Carbonneau et al. (2005); Farò et al. (2025); Lang et al. (2020).

Recent studies have demonstrated that integrating optical imagery with lidar-derived topographic predictors can substantially improve classification performance by incorporating direct measures of surface roughness Chardon et al. (2020); Rogers et al. (2025). These multimodal approaches highlight the importance of combining spectral and structural information for reliable substrate discrimination, but they also introduce significant data acquisition costs and operational complexity, limiting their applicability in routine UAV-based monitoring.

In addition to these methodological challenges, a critical and often overlooked limitation in substrate classification workflows is the influence of spatial autocorrelation (SAC), whereby neighboring observations exhibit statistical dependence Tobler (1970). Because both RGB-based and multimodal approaches rely on spatially structured data, conventional validation strategies that randomly partition samples do not ensure independence between training and testing sets. This leads to data leakage (i.e., shared spatial information between training and validation samples) and artificially inflated accuracy estimates Meyer et al. (2019); Roberts et al. (2017); Wadoux et al. (2021). This issue is particularly pronounced in high-resolution orthophotos, where spectral and structural patterns vary smoothly across space. Despite its importance, SAC is rarely addressed explicitly in fluvial substrate classification studies, and computationally efficient validation strategies that account for spatial dependence remain underdeveloped.

This study addresses these methodological and validation gaps by proposing a scalable workflow for categorical substrate classification using ultra-high-resolution ("hyperspatial", sensu Piégay et al. (2020); resolution < 10 cm) UAV-derived RGB imagery. The proposed framework integrates spectral color information with statistical texture descriptors derived from Gray-Level Co-occurrence Matrices (GLCM) and structural features extracted using Local Binary Patterns (LBP), enabling reliable discrimination of substrate classes without requiring lidar-derived predictors Haralick (1979); Ojala et al. (2002). The framework is ad-hoc developed using open-source tools and standard GIS

 doi:10.20944/preprints202605.1410.v1

platforms, avoiding reliance on proprietary software and reducing barriers to adoption in operational and research contexts.

As the framework relies exclusively on optical imagery, its applicability is inherently constrained by water clarity. Direct visibility of the riverbed is therefore a prerequisite, limiting the method to conditions of low turbidity where the water column is optically transparent and free of significant suspended sediment Carbonneau et al. (2004a).

In parallel, the study explicitly evaluates the impact of spatial autocorrelation on model performance by comparing standard and spatially aware cross-validation strategies, and by implementing a design-based probability sampling approach for independent accuracy assessment Wadoux et al. (2021). The workflow is validated across four geomorphologically and hydraulically distinct river reaches, demonstrating both high classification performance and rigurous validation.

2. Methodology

The methodology is structured as a semi-automated, five-stage workflow (Figure 1), designed for independent execution at each study site to produce a locally tuned classification model. The workflow comprises: (1) expert-based ground-truth substrate annotation; (2) feature set generation; (3) spatially aware model optimization; (4) full-domain classification; and (5) design-based validation for independent accuracy assessment.

The complete workflow is implemented as an open-source Jupyter Notebook framework and is publicly available at https://github.com/tsotop/fluvialsubstrate.

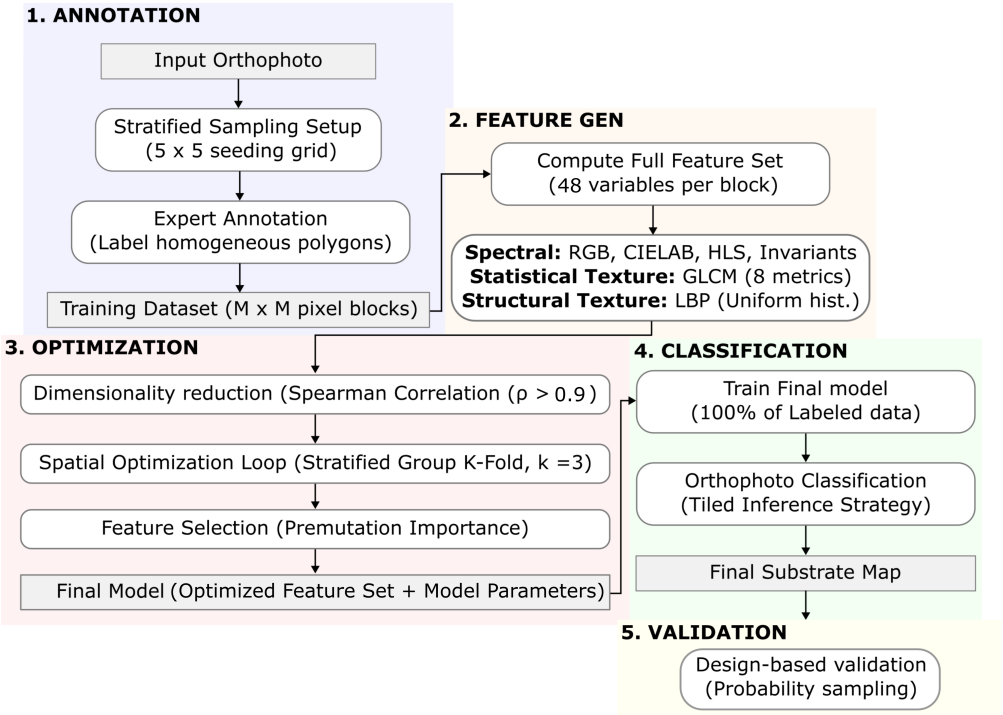

Figure 1. Conceptual overview of the five-stage classification workflow. The pipeline progresses from site-specific data annotation (1) through feature engineering (2) and spatially-aware optimization (3) to full-domain classification (4), followed by design-based validation (5) for independent accuracy assessment. Gray-shaded boxes indicate the output of each main step.

2.1. Ground Truth Annotation

Ground truth data are established through the manual delineation of vector polygons over areas exhibiting a visually homogeneous and dominant substrate class, as interpreted from high-resolution RGB orthophotos. This process is conducted entirely in a desktop environment (e.g. QGIS), consistent with remote sensing classification workflows where access to field data is constrained or where image resolution allows direct expert interpretation Foody (2002).

 doi:10.20944/preprints202605.1410.v1

To initiate the annotation process, annotators perform an exploratory visual scan of the orthophoto to identify the full range of substrate types, thereby establishing an informal class dictionary. Defining these categories a priori ensures that class annotations are applied in a mutually exclusive and semantically consistent manner, which is a fundamental requirement for accurate classification Foody (2002). Substrate classes are defined with sensitivity to radiometric differences; in particular, wet and dry variants of the same material are annotated as separate classes to account for spectral shifts induced by moisture conditions Carbonneau et al. (2004a). This distinction improves class separability and reduces ambiguity in the feature space.

To guide the spatial distribution of samples and mitigate geographic bias, the tool partitions the target domain into a 5×5 grid. Within each cell, five candidate points are seeded at random, ensuring spatial coverage across the domain. This design is particularly important in riverine environments, where substrate classes vary systematically due to longitudinal, lateral and vertical sorting processes Knighton (2014); Powell (1998); Stehman (2012). A purely random scheme may under-represent specific geomorphic sub-domains, whereas stratification ensures that all regions of the orthophoto are sampled. The seeded points are intended to ensure spatial coverage rather than class balance. Annotators are therefore encouraged to supplement these points with additional polygons to adequately represent all substrate classes identified in the class dictionary, particularly for less frequent or spatially limited classes. As a practical guideline, a minimum of three annotated polygons per class is recommended to support the spatial cross-validation framework and ensure robust model training. These seeded points act as spatial prompts rather than strict sampling locations. If a point intersects a transitional zone (e.g., water boundaries or mixed substrates), the user may annotate a nearby homogeneous polygon that best represents a single substrate class.

Finally, the fundamental unit of analysis is defined as a non-overlapping $M \times M$ pixel block, corresponding to a fixed physical area determined by image resolution (e.g., $M = 50$ for a 2 cm orthophoto yields a 1×1 m block). Feature values are aggregated within each block, reducing pixel-level noise and enabling robust extraction of texture descriptors that require spatial context Carbonneau et al. (2004b); Woodget and Austrums (2017).

The size of the block M is configurable but is standardized to approximately 1 m^2 for this study, consistent with the standard protocols used in the mapping of river habitat on the mesoscale Parasiewicz (2007); Vezza et al. (2014). This scale is sufficiently large to capture representative textural information while remaining small enough to resolve fine-scale spatial heterogeneity across the study domain.

2.2. Feature Set Generation

A comprehensive feature set (48 parameters) is systematically computed for each analysis block. This set serves as the candidate pool for subsequent model optimization and is grouped into three distinct feature families.

To provide an overview of the feature space, Table 1 summarizes the set of predictors computed for each analysis block, grouped by feature family. For clarity, representative variables are listed, while features computed across multiple channels or bins (e.g., GLCM metrics and LBP histograms) are reported in aggregated form.

 doi:10.20944/preprints202605.1410.v1

Table 1. Summary of computed feature set grouped by family.

Type	Name	Description
Spectral & First-Order Features		
Spectral	R, G, B	Mean RGB intensity
Spectral	μ_{L^*}	CIELAB lightness
Spectral	μ_L (HLS)	HLS luminance
Spectral	μ_{a^*}/μ_{L^*}	Norm. green–red
Spectral	μ_{b^*}/μ_{L^*}	Norm. blue–yellow
Spectral	c_1, c_2, c_3	Color invariants
First-order	σ_{L^*}, σ_L	Std. dev. brightness
First-order	$\mathrm{Var}_{L^*}, \mathrm{Var}_L$	Brightness variance
Statistical Texture (GLCM)		
GLCM	Contrast	Intensity contrast
GLCM	Dissimilarity	Local variation
GLCM	Homogeneity	Uniformity
GLCM	ASM	Energy (squared sum)
GLCM	Energy	Texture uniformity
GLCM	Correlation	Linear dependency
GLCM	Entropy	Randomness
GLCM	Neg. Entropy	Inverse entropy
Structural Texture (LBP)		
LBP	$\mathrm{LBP}_{P,R}$	Local binary pattern
LBP	Uniform patterns	Rotation invariant
LBP	Histogram (10 bins)	Texture distribution

Note: GLCM metrics and LBP histograms are computed independently on both CIELAB Lightness (L^*) and Luminiosity (HLS-L) channels.

First, spectral and first-order features are computed to capture general radiometric properties. These include the mean intensity of the red, green, and blue channels (R, G, B), directly derived from the RGB orthophoto. Color space transformations are applied to obtain the mean lightness component from the CIELAB color space (μ_{L^*}), representing perceptual brightness, and the mean luminance from the HLS color space (μ_L), corresponding to intensity in cylindrical color representation, both computed at the block level.

Beyond spectral intensity, normalized color ratios (μ_{a^*}/μ_{L^*} and μ_{b^*}/μ_{L^*}) are derived from CIELAB, where μ_{a^*} and μ_{b^*} represent chromatic components (green–red and blue–yellow axes). This normalization reduces sensitivity to illumination variability and improves facies discrimination Rogers et al. (2025). Additionally, color invariants (c_1, c_2, c_3) are computed to enhance robustness to shadowing Gevers and Smeulders (1999).First-order statistics of the L^* and HLS-L channels, including standard deviation and variance, are computed to characterize sub-block radiometric variability.

Second, statistical texture features are derived using the Gray-Level Co-occurrence Matrix (GLCM). For each analysis block, grayscale representations of the L^* and HLS-L channels are converted to 8-bit intensity and used to compute a normalized, symmetric GLCM with pixel offset distance $d = 1$ and orientation $\theta = 0^\circ$ (horizontal direction). The number of gray levels is adaptively determined from the intensity range within each block (up to 256 levels). From each GLCM, eight standard metrics (Contrast, Dissimilarity, Homogeneity, Angular Second Moment-ASM, Energy, Correlation, Entropy, Negative Entropy) are extracted Carbonneau et al. (2005); Woodget and Austrums (2017).

GLCM features are computed directly at the block scale from pixel values contained within each block, without incorporating information from neighboring blocks. For boundary blocks intersecting annotation edges, masked pixels outside the polygon footprint may be included in the rectangular patch.

Third, structural texture features are computed using Local Binary Patterns (LBP) to capture local micro-structures such as edges and corners. LBP codes are computed at the pixel level using the

 doi:10.20944/preprints202605.1410.v1

uniform operator with $P = 8$ neighbors and radius $R = 1$, and aggregated within each block into a normalized histogram (10 bins) to form block-level descriptors Ojala et al. (2002); Pietikäinen and Zhao (2015).

Features are computed independently for the L^* and HLS-L channels using only pixels contained within each block, with no spatial interaction across block boundaries. Boundary blocks may include masked pixels outside annotated regions, which can influence the resulting histograms.

2.3. Feature Selection and Optimization

A Random Forest (RF) classifier is selected as the core algorithm due to its strong performance in high-dimensional and multicollinear feature spaces typical of object-based image analysis Belgiu and Drăguţ (2016). A standard ensemble configuration is adopted (100 trees, no maximum depth constraint).

To construct a robust and computationally efficient classifier, a structured feature selection workflow is applied. First, to reduce multicollinearity, pairwise Spearman rank correlations are computed across all features. For feature pairs exceeding a correlation threshold ($|\rho_s| > 0.9$), one variable is removed by retaining the earlier-ordered feature in the list and pruning the later one. This deterministic filtering reduces redundancy while preserving a stable and interpretable feature set.

Second, feature relevance is assessed using cross-validated Permutation Feature Importance (PFI), which quantifies the decrease in model performance when a feature is randomly permuted Strobl et al. (2007). For each cross-validation fold, the model is trained on the corresponding training subset and evaluated on the held-out data using accuracy. Each feature is then permuted individually on the validation subset, and the resulting decrease in accuracy is recorded. This permutation is repeated five times per feature within each fold. Consequently, each feature yields $5 \times K$ importance estimates (with K folds), which are aggregated as mean (μ) and standard deviation (σ) to quantify both effect size and stability across folds.

Feature pruning is then performed using a conservative uncertainty-aware criterion based on permutation importance. Features are removed only when the upper bound of their importance estimate remains non-positive (i.e., $\mu + \sigma \leq 0$), ensuring that predictors are excluded only when they consistently fail to contribute to model performance. This approach accounts for the known variability of permutation importance estimates in ensemble models, particularly in the presence of correlated predictors and limited sample sizes Breiman (2001); Fisher et al. (2019). Beyond improving model interpretability and stability, feature pruning also reduces computational cost, as only the selected subset of predictors needs to be computed during full-domain classification.

Finally, model optimization and feature evaluation are conducted within a spatially aware cross-validation framework using Stratified Group K-Fold (SGKF). In K-Fold cross-validation, the dataset is partitioned into k subsets (folds), where each fold is iteratively used as a testing set while the remaining folds are used for training, ensuring that all samples are evaluated. In this implementation, grouping is defined by polygon identifiers to enforce spatial independence between training and testing samples, thereby mitigating data leakage (i.e., the unintended sharing of spatially correlated information between training and testing samples) due to spatial autocorrelation. Stratification ensures that class proportions are preserved across folds Schratz et al. (2019); Wadoux et al. (2019). The number of folds (k) is constrained to match the number of polygons in the least represented class, ensuring that each fold contains at least one instance of every class. To balance robustness and sample sufficiency, k is bounded between 3 (minimum recommended number of annotations per class) and 10 folds.

2.4. Final Model Training and Full-Domain Classification

Upon completion of the optimization phase, the final model configuration is formalized. The final classifier is trained using the complete annotated dataset for each study site, thus maximizing predictive performance by leveraging the full range of observed surface sediment substrate variability. Class imbalance is addressed through the use of class-weighted training, where weights are defined

 doi:10.20944/preprints202605.1410.v1

inversely proportional to class frequencies to improve the recall of underrepresented substrates Colditz (2015).

Subsequently, a spatially continuous substrate map is generated for the entire orthophoto. To manage memory constraints associated with high-resolution imagery, a tiled inference strategy is employed Ma et al. (2015). The orthophoto was partitioned into non-overlapping tiles of 1000×1000 pixels, which are processed sequentially. Within each tile, the image is subdivided into non-overlapping $M \times M$ analysis blocks, and only the selected subset of features, determined in the previous step, is computed for each block. The trained model)ethen assigns a substrate class to each block, and the results are mosaicked to produce the final classified raster.

2.5. Design-based Validation

While the cross-validation described above assesses model consistency during training, it does not directly quantify the accuracy of the final classified map across the full spatial domain. As demonstrated by Wadoux et al. (2021), operational map accuracy typically lies between the optimistic estimates obtained from standard cross-validation and the more conservative estimates derived from spatial cross-validation. Here, standard cross-validation refers to random partitioning of samples irrespective of spatial location, whereas spatial cross-validation enforces spatial separation between training and testing samples through polygon-based grouping.

To provide a statistically grounded estimate of map accuracy, a probability sampling approach is proposed as a final validation step. Following the design-based inference framework of Wadoux et al. (2021), validation samples are generated using stratified random sampling based on the classified map. For each mapped class h, the number of validation points (n_h) is initially allocated proportionally to its mapped area (W_h), such that $n_h \propto W_h \cdot N$, where N is the target total sample size (set to 150 by default). To ensure adequate representation of rare classes, a minimum sample size constraint is enforced by assigning each class at least $n_{\min} = 20$ samples, such that the final allocation for each class corresponds to the larger of the proportional allocation or the minimum threshold.

In cases where the number of available blocks within a class is lower than the allocated sample size, all blocks are selected. Consequently, the total number of validation samples is not fixed a priori and could deviate from the target value depending on class distribution and minimum sampling constraints.

All validation blocks are then independently annotated by an expert interpreter through visual inspection within the GIS environment. Under this framework, map accuracy is treated as a population parameter and estimated from a probability sample in which each spatial unit has a known, non-zero inclusion probability, thereby enabling unbiased estimation of overall and class-specific accuracies.

Final accuracy metrics are computed using design-based inference estimators, where each validation sample contributes proportionally to the mapped area of its corresponding class. For each class h, the sample-based User's Accuracy (p_h) is estimated as:

$$p_h = \frac{n_{h,\text{correct}}}{n_h} \tag{1}$$

where n_h is the number of validation samples within class h, and $n_{h,\text{correct}}$ is the number of correctly classified samples in that stratum.

The Overall Accuracy ($\hat{O}$) is then estimated as a weighted sum of per-class accuracies:

$$\hat{O} = \sum_{h=1}^{K} W_h \cdot p_h \tag{2}$$

where W_h is the proportion of the total mapped area occupied by class h, ensuring that each class contributes proportionally to its spatial extent.

Uncertainty metrics, including standard error and confidence intervals, are also derived using standard design-based estimators.

 doi:10.20944/preprints202605.1410.v1

3. Experimental Applications

To evaluate the robustness and transferability of the proposed workflow, the methodology is applied to four distinct gravel-bed river reaches in North East Italy (Figure 2). These sites are selected to capture a representative spectrum of channel morphologies, ranging from single-thread channelized to meandering, wandering and braided systems, thereby allowing testing the model's applicability across a wide spectrum of channel morphologies and their corresponding surface sedimentary patterns.

Orthophotos were generated from UAV-acquired RGB imagery using standard Structure-from-Motion (SfM) photogrammetry workflows. Image processing and orthomosaic generation were performed using commercial software (e.g., Agisoft Metashape). All datasets were georeferenced using ground control points (GCPs), ensuring consistent spatial accuracy across sites.

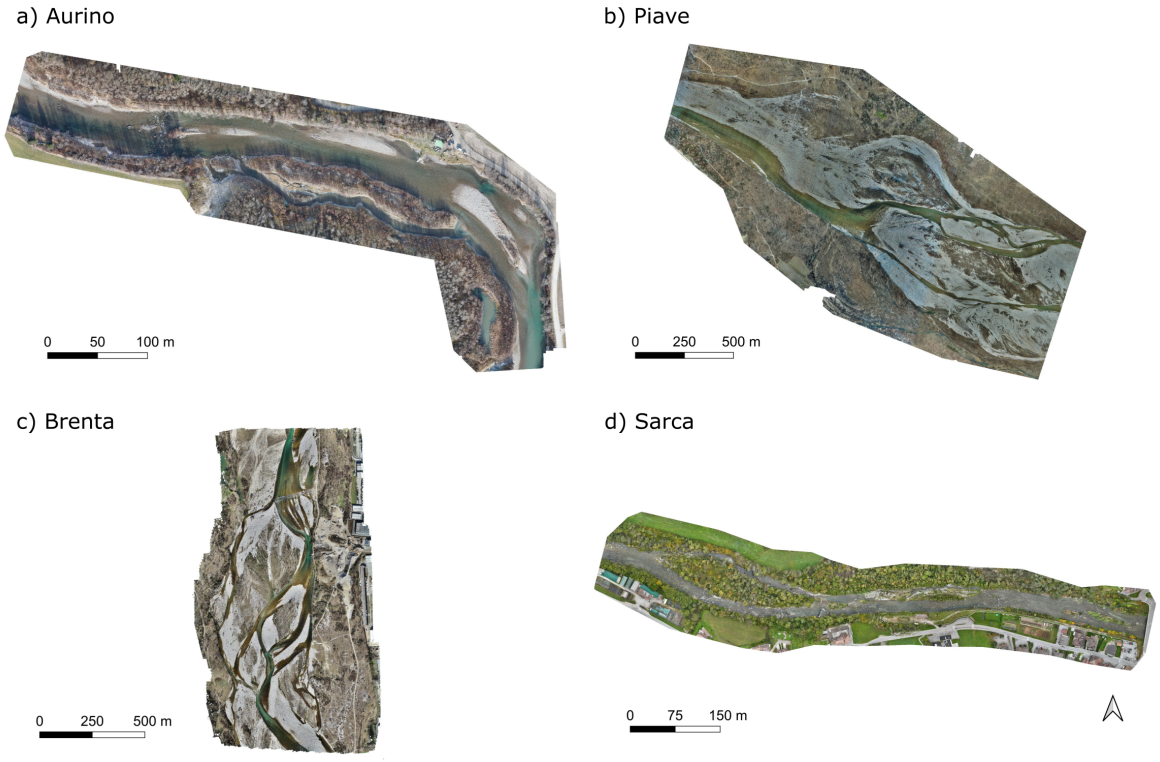

Figure 2. Orthophoto of the four study river reaches: (a) Aurino, (b) Piave, (c) Brenta, (d) Sarca.

Table 2 summarizes the key acquisition and geomorphic parameters for each site. The ground sampling distance (GSD) varies slightly across datasets (1.5 to 2.0 cm/px).

Table 2. Summary of key acquisition and geomorphic characteristics for the four experimental study sites.

Study Site	Channel morphology	UAV Platform	GSD (cm/px)	Domain Area (ha)	Substrate Range
Aurino	Meandering	DJI S-1000	1.5	18.06	Sand to Boulders
Piave	Wandering	DJI S-1000	2.0	368.17	Sand to Large Cobble
Brenta	Braided	DJI Phantom 4 RTK	2.0	149.46	Sand to Large Cobble
Sarca	Channelized, single-thread	DJI Mavic 3M	2.1	24.72	Gravel to Bedrock

4. Results

4.1. Ground Truth Generation and Dataset Characteristics

The semi-automated annotation workflow was applied to all four study sites. The initial stratified grid seeding generated 125 candidate points per orthophoto; points falling outside the active river channel were automatically discarded, resulting in 40–50 valid seeding locations per site (Figure 3). These points guided the expert delineation of homogeneous training polygons.

 doi:10.20944/preprints202605.1410.v1

Table 3 summarizes the resulting datasets. The number of distinct polygons ranged from 37 (Sarca) to 55 (Brenta). The total volume of training data varied significantly, from 6,774 blocks in the Sarca reach to nearly 58,000 blocks in the Piave, primarily driven by the differing spatial extent of the active channel. All datasets included wet/dry class distinctions (e.g., `wet-cobble` vs. `dry-cobble`). Computationally, the feature extraction pipeline required between 14 seconds (Sarca) and 151 seconds (Piave) on an Apple M4 Pro workstation with 48GB RAM.

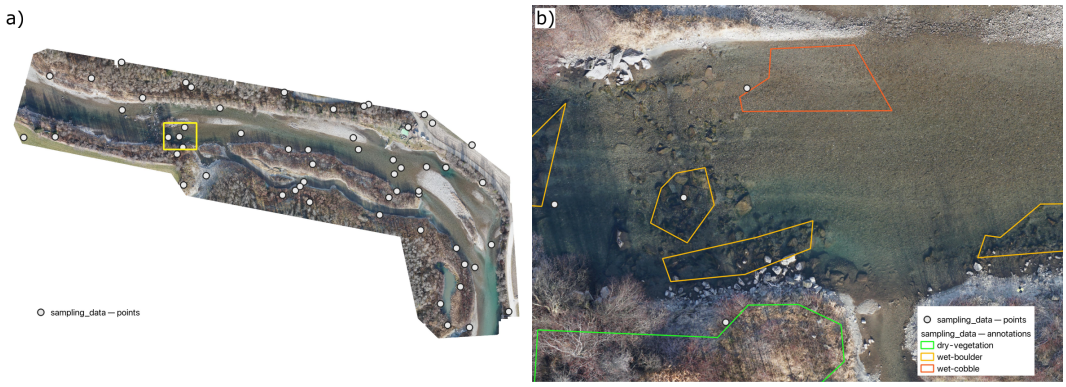

Figure 3. Visualization of the ground truth generation workflow applied to the Aurino study site. (a) Full domain orthophoto overlaid with the stratified random sampling points (white dots). The yellow box indicates the extent of the detail view. (b) Zoomed-in view of the highlighted region showing the final expert-delineated training polygons. Polygons are colored by substrate class (see legend).

Table 3. Summary of the ground-truth training datasets generated for each study site, including the detailed breakdown of blocks per substrate class. Substrate classes are defined by characteristic grain-size ranges: Sand (0.06–2 mm), Cobble Small (6–20 cm), Cobble Large (20–40 cm), and Boulder (>40 cm). Bedrock represents continuous exposed rock substrate, rather than unconsolidated sediment particles.

Study Site	Classes	Polygons	Total Blocks	Processing Time (s)	Class Breakdown (Number of Blocks)
Aurino	7	51	8,591	20	Dry Vegetation (3565), Wet Cobble Large (2337), Dry Cobble Large (1195), Wet Sand (563), Wet Boulder (550), Dry Sand (248), Dry Boulder (133)
Piave	6	53	57,981	151	Dry Vegetation (33082), Wet Cobble Large (8253), Wet Cobble Small (6394), Dry Cobble Large (4972), Dry Cobble Small (3028), Dry Sand (2252)
Brenta	6	55	11,908	24	Dry Vegetation (4061), Wet Cobble Large (2201), Dry Cobble Small (1869), Wet Cobble Small (1798), Dry Cobble Large, (1164), Dry Sand (815)
Sarca	6	37	6,774	14	Dry Vegetation (4227), Wet Bedrock (961), Wet Boulder (837) Wet Cobble Large (464), Dry Boulder (173), Dry Cobble Large (112)

4.2. Feature Selection and Model Optimization

The Permutation Feature Importance (PFI) analysis identified consistent patterns of predictive variables across the four river reaches (Figure 4). Prior to ranking, a Spearman's correlation filter ($\rho > 0.9$) was applied to remove highly collinear predictors from the initial 48-variable feature set. The final feature space was reduced to 22 variables for Aurino, 23 for Piave, 14 for Brenta, and 19 for Sarca.

 doi:10.20944/preprints202605.1410.v1

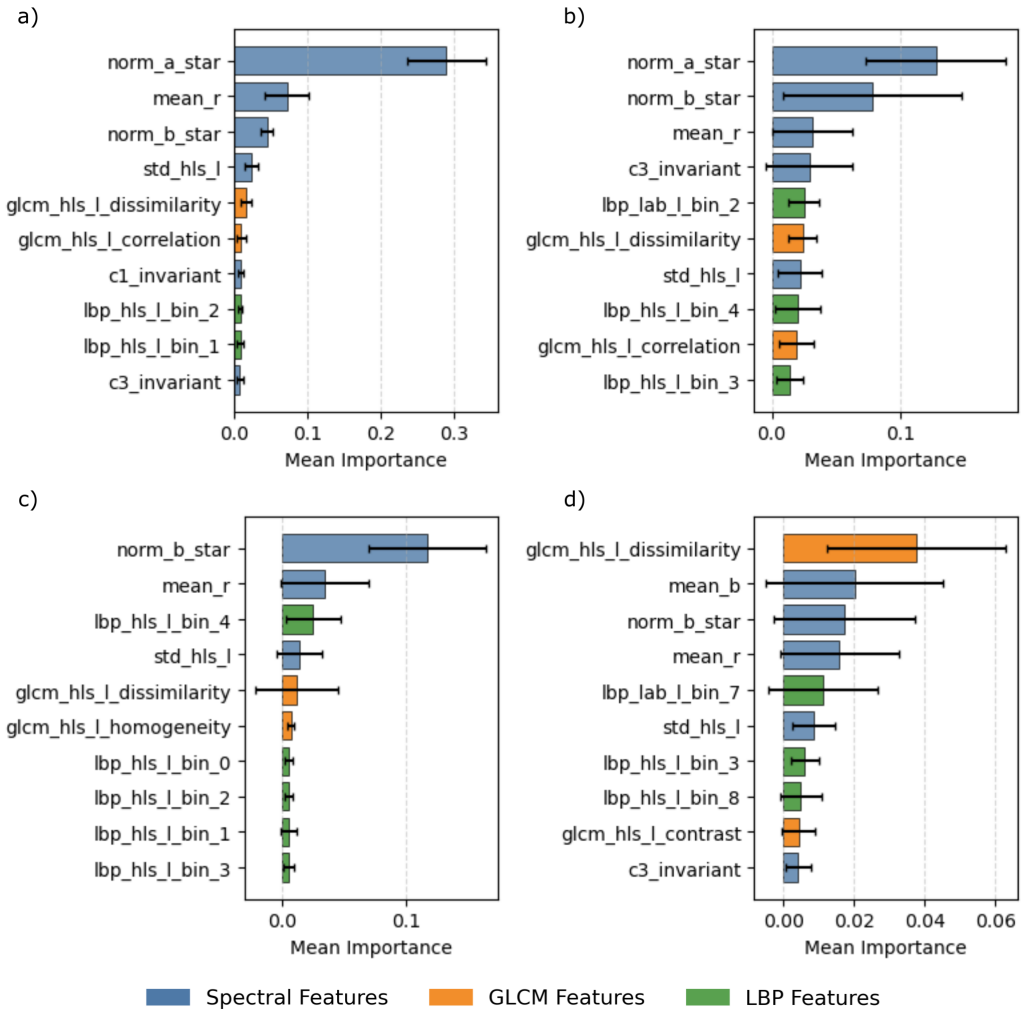

Figure 4. Ranked feature importance for the four study sites: (a) Aurino, (b) Piave, (c) Brenta, and (d) Sarca. Horizontal bars display the top 10 most predictive variables identified by the spatial optimization loop. Features are colored by family (Blue: Spectral; Orange: GLCM; Green: LBP). `norm_a_star` and `norm_b_star` refer to the normalized CIELAB chromaticity ratios; c1-c3 invariant refer to the shadow-robust color invariants; glcm [metric] denotes statistical texture features derived from the Grey-Level Co-occurrence Matrix; and lbp [channel] bin [n] refers to specific bins (0–9) of the Local Binary Pattern histogram computed on either the L* or HLS-L channel. The complete feature list is provided in Supplementary Table S1

Across all case studies, spectral features ranked highest. Normalized color ratios (`norm_a_star`, `norm_b_star`) and mean spectral intensities (e.g., `mean_r`, `mean_b`) consistently occupied the top positions.

GLCM-based texture features, particularly `dissimilarity`, were present among the top-ranked predictors in all datasets and ranked first in the Sarca reach (Figure 4d). LBP features were also represented within the top 10 predictors across all case studies, with multiple LBP variables appearing in the Brenta and Sarca datasets.

4.3. Model Training and Preliminary Performance Assessment

Table 4 presents the preliminary model performance metrics obtained from Standard and Spatial Cross-Validation. Standard Cross-Validation provides an optimistic estimate of performance because samples are partitioned without accounting for spatial proximity, whereas Spatial Cross-Validation provides a more conservative assessment by enforcing spatial separation between training and testing samples. The difference between these two estimates is interpreted here as the "optimism gap",

 doi:10.20944/preprints202605.1410.v1

representing the inflation of model performance caused by spatial dependence between training and testing samples.

It is important to note that cross-validation performance is evaluated using the F1-score and Overall Accuracy (OA), whereas final map accuracy is assessed separately using design-based overall accuracy, which reflects the area-weighted reliability of the classified map.

Standard Cross-Validation consistently produced the highest F1 estimates, exceeding 0.8 across all four sites. When accounting for spatial autocorrelation using Spatial Cross-Validation (SGKF), mean F1 estimates decreased to between 0.59 and 0.64. Across all case studies, the difference between Standard and Spatial F1 ranged from approximately 20 to 32 percentage points.

Table 4. Comparison of model performance across validation strategies. Metrics include F1-score and Overall Accuracy (OA) for Standard and Spatial Cross-Validation. The "optimism gap" refers to the inflation of performance estimates under Standard Cross-Validation due to spatial autocorrelation.

Study Site	Standard CV (F1/OA)	Spatial CV (F1/OA)
Aurino	0.85/0.95	0.59/0.80
Piave	0.96/0.98	0.64/0.82
Brenta	0.94/0.95	0.59/0.64
Sarca	0.80/0.92	0.61/0.84

4.4. Full-Domain Classification and Final Map Validation

The optimized models are applied to generate spatially continuous substrate maps for the full extent of each orthophoto. Representative results for the Aurino and Piave sites are shown in Figure 5, while the Brenta and Sarca maps are provided in the Supplementary Material (Figure S1).

The Design-based Overall Accuracy, calculated through the design-based inference on the final classified raster, yielded values of 83.4% (Aurino), 80.5% (Piave), 69.8% (Brenta), and 95.2% (Sarca). Notably, for the Sarca dataset, the final map accuracy exceeded the mean Spatial CV estimate by 10.9 percentage points (Table 4).

Table 5 reports class-specific User's Accuracy (UA) for each study site. Across all sites, dominant and well-represented classes such as `Dry Sand` and `Dry Vegetation` consistently achieved high accuracies ($>$ 80%). In contrast, classes with limited spatial representation exhibited greater variability in performance. For example, in the Brenta dataset, `Wet Cobble (Small)` (1,798/11,908 total blocks) achieved a relatively low accuracy (31.58%), while in the Sarca dataset, `Dry Cobble (Large)` (112/6,774 total blocks) recorded 0% User's Accuracy.

The lower performance observed in the Brenta reach is likely related to the presence of multiple adjacent substrate classes along a gradual grain-size continuum, rather than clearly separated end-member classes. This increases confusion between visually similar categories, particularly among cobble-size classes. Overall, these variations highlight the combined influence of class imbalance, semantic overlap between substrate categories, and limited sample sizes within individual strata under the design-based estimation framework.

 doi:10.20944/preprints202605.1410.v1

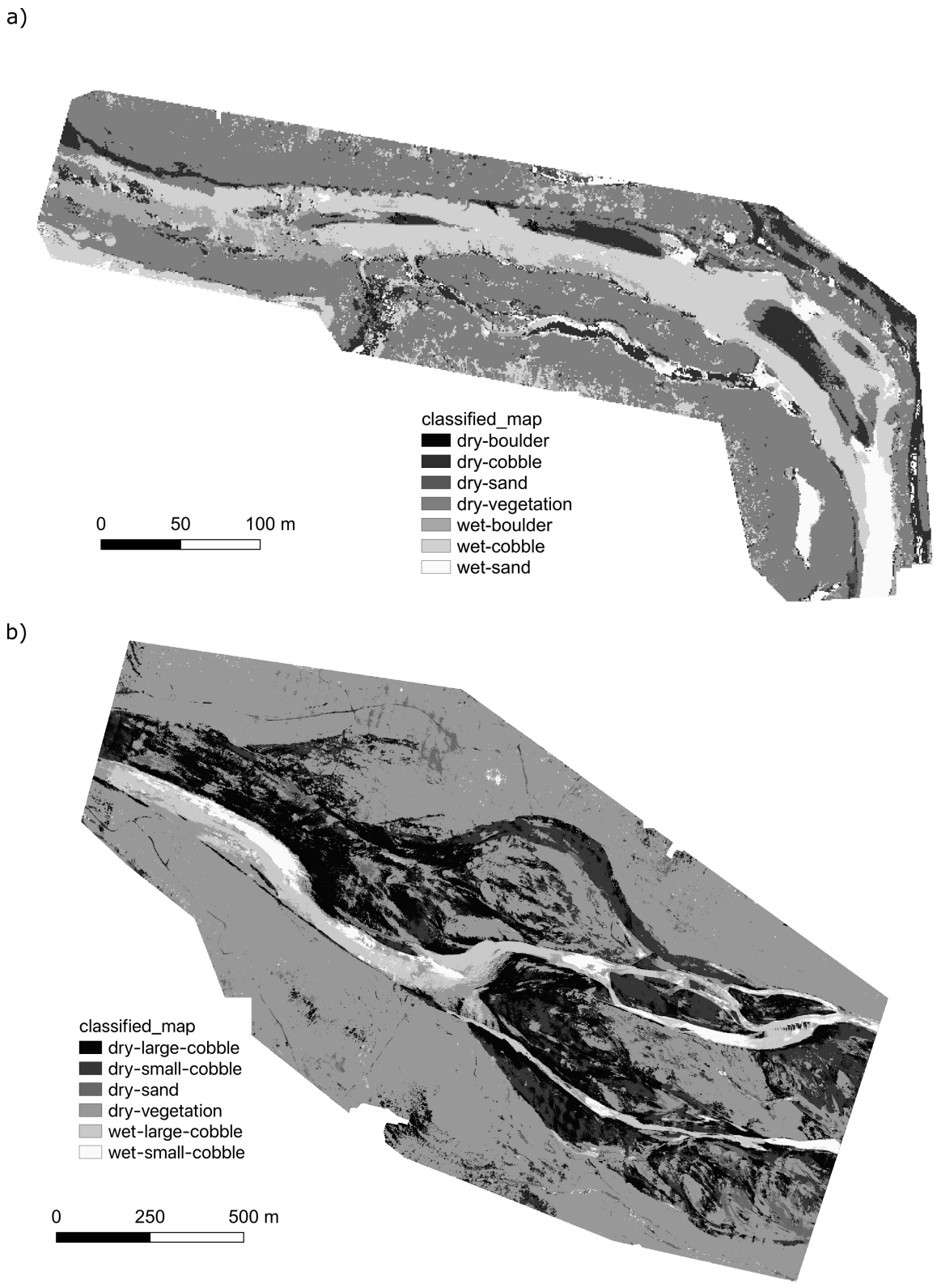

Figure 5. Classified substrate maps for (a) Aurino and (b) Piave study sites. Each map shows the spatial distribution of substrate classes predicted by the optimized Random Forest model at the block scale.

 doi:10.20944/preprints202605.1410.v1

Table 5. Detailed User's Accuracy per substrate class derived from design-based probability sampling. Dashes indicate classes not present in a specific study site.

	User's Accuracy by Study Site			
Substrate Class	**Aurino**	**Piave**	**Brenta**	**Sarca**
Dry Boulder	—	—	—	81.82%
Dry Cobble (Large)	72.22%	51.85%	57.89%	0%
Dry Cobble (Small)	46.67%	85.00%	83.33%	—
Dry Sand	91.67%	83.33%	90.91%	—
Dry Vegetation	89.19%	87.36%	63.46%	100.00%
Wet Bedrock	—	—	—	100.00%
Wet Boulder	—	—	—	89.47%
Wet Cobble (Large)	93.10%	95%	61.11%	55.56%
Wet Cobble (Small)	88.89%	73.68%	31.58%	—
Estimated Overall Accuracy	**83.42%**	**80.53%**	**69.77%**	**95.21%**

5. Discussion

This study presents a scalable and spatially robust framework for riverine substrate (sediment size) classification using ultra-high-resolution RGB imagery alone. By integrating spectral and structural texture features within a rigorous validation framework, the proposed methodology addresses key limitations of existing approaches, including reliance on multimodal data, and optimistic performance reporting. The results demonstrate that accurate and transferable substrate classification can be achieved without lidar-derived predictors, provided that spatial structure is explicitly accounted for and that feature selection captures both radiometric and textural variability.

5.1. Feature Complementarity and the Texture–Spectral Trade-off

Feature importance analysis reveals a consistent hierarchy of predictors across all study sites, with spectral features providing the primary discriminatory signal. In particular, normalized color ratios and mean spectral intensities ranked highest in the models, indicating that differences in spectral response are key for separating substrate classes in this study Carbonneau et al. (2004a). This also supports the distinction between wet and dry substrate classes, which exhibit consistent radiometric differences in the imagery. These findings are consistent with Rogers et al. (2025), who identified color and color variability as important predictors for distinguishing substrate facies.

Statistical texture features derived from the Gray-Level Co-occurrence Matrix (GLCM), particularly dissimilarity, were consistently among the most important predictors and emerged as the top-ranked feature in the Sarca dataset. GLCM-based metrics have long been established as effective descriptors of sediment texture in fluvial remote sensing, forming the basis of early image-based grain size estimation approaches Carbonneau et al. (2004b); Verdú et al. (2005). The relatively large variability associated with dissimilarity in this study suggests sensitivity to local spatial heterogeneity, highlighting the influence of small-scale structural variability in complex river environments.

Structural texture features derived from Local Binary Patterns (LBP) were also consistently present among the top predictors, contributing additional discriminatory power by capturing local micro-structural patterns such as grain edges and surface roughness. Their increased presence in the Brenta and Sarca datasets indicates a greater reliance on structural descriptors in settings where spectral separability is reduced.

Overall, these results demonstrate a clear complementarity between spectral and texture-based predictors. While spectral features capture first-order radiometric differences, texture descriptors

 doi:10.20944/preprints202605.1410.v1

enhance class separability by representing spatial structure and sediment organization. This combined feature space enables robust classification across a range of geomorphic conditions.

Importantly, the ability of image-derived texture features to capture structural variability represents a key advantage over purely spectral approaches. Similar information is often obtained from lidar-derived roughness metrics Rogers et al. (2025); Woodget and Austrums (2017), but the present results demonstrate that structural characteristics can be inferred from high-resolution RGB imagery alone.

5.2. The "Optimism Gap" and the Necessity of Spatial Validation

Our results highlight that standard non-spatial validation methods tend to produce optimistic performance estimates in riverine imagery. Across all case studies, standard cross-validation yielded F1 scores exceeding 0.80, representing an upper-bound estimate of model performance. However, when spatial autocorrelation was explicitly controlled using the Stratified Group K-Fold (SGKF) strategy, performance decreased substantially (by 19–35% across sites), reflecting the true model's ability to generalize to spatially independent data.

This discrepancy arises from spatial autocorrelation (SAC), whereby neighboring observations share similar spectral and structural characteristics Tobler (1970). As a result, randomly partitioned validation data violate the assumption of independence, leading to data leakage and inflated accuracy estimates Meyer et al. (2019); Roberts et al. (2017); Wadoux et al. (2021).

Importantly, spatial cross-validation should not be interpreted as a pessimistic estimate of final performance, but rather as a conservative test of model robustness under spatial independence. Its primary role in this workflow is to guide model optimization while preventing overfitting to spatially clustered training data.

To obtain an unbiased estimate of map accuracy, a design-based validation framework was implemented as a final, independent assessment. Unlike cross-validation, which evaluates model performance on subsets of the training data, design-based inference treats accuracy as a population parameter and estimates it from a probability sample of the fully classified map Wadoux et al. (2021).

This pattern is consistent across all study sites. For Aurino, Piave, and Brenta, the design-based Overall Accuracy (83.42%, 80.53%, and 69.77%, respectively) closely matches or exceeds the corresponding spatial cross-validation estimates (80%, 82%, and 64%).

In contrast, the Sarca dataset exhibits a larger discrepancy, with the final design-based accuracy (95.21%) exceeding the spatial cross-validation estimate (84%). This indicates that while spatial cross-validation provides a conservative estimate of model performance, the final model—trained on the full dataset—can achieve higher operational accuracy once all available information is incorporated.

Together, these results are consistent with previous findings Rogers et al. (2025); Wadoux et al. (2021) and highlight the importance of spatially explicit validation in preventing overly optimistic performance estimates. At the same time, design-based validation provides an independent and statistically unbiased estimate of final map accuracy, reflecting operational model performance. In this context, spatial cross-validation should be interpreted as a conservative assessment of model generalization, particularly in the presence of spatial autocorrelation and class imbalance, where the partitioning of rare classes across folds can lead to reduced performance estimates. By contrast, design-based validation evaluates the fully trained model over the mapped domain, thereby avoiding both optimistic bias and the conservative effects introduced during spatial partitioning. Together, these approaches provide a more complete and reliable characterization of model performance than any single validation strategy alone.

5.3. Comparison with Existing RGB-Based Substrate Mapping Approaches

The proposed framework differs from existing RGB-based substrate classification methods in several key aspects, including validation strategy, feature representation, and operational accessibility.

From a feature representation perspective, the proposed framework extends traditional RGB-based approaches by integrating structural texture descriptors through Local Binary Patterns (LBP).

 doi:10.20944/preprints202605.1410.v1

While previous studies primarily rely on spectral information and second-order statistical textures (GLCM) Arif et al. (2017); Carbonneau et al. (2005), LBP features were consistently present among the top-ranked predictors across all case studies. Their increased representation in certain datasets suggests that structural texture provides complementary information to spectral and GLCM features, particularly in environments where spectral contrast alone may be insufficient.

A primary distinction lies in the explicit treatment of spatial autocorrelation (SAC) during model validation. Previous studies, such as Arif et al. (2017) and Giroux et al. (2022), rely on random sampling strategies that do not enforce spatial independence between training and validation data. In the case of Giroux et al. (2022), validation points were selected within training polygons, a practice known to introduce data leakage and inflate accuracy estimates Misiuk et al. (2021); Roberts et al. (2017). Similarly, Arif et al. (2017) employed random sampling without accounting for spatial dependence. In contrast, the present study explicitly quantifies the resulting "optimism gap" by comparing standard cross-validation with a Stratified Group K-Fold (SGKF) approach, which enforces spatial independence by grouping samples at the polygon level. This is further complemented by design-based probability sampling, providing a statistically unbiased estimate of map accuracy.

Another key difference lies in operational accessibility. Existing workflows often depend on complex and fragmented software environments. For instance, Arif et al. (2017) utilized a combination of commercial software, including ERDAS IMAGINE, MATLAB, and eCognition, while Giroux et al. (2022) relied on ArcGIS and PCI Geomatica. In contrast, the proposed method is implemented as an open-source, Python-based workflow in combination with standard/open GIS platforms (QGIS in this implementation). The process is reduced to two main steps: manual annotation and automated classification, making it accessible to practitioners without specialized expertise in computer vision or access to commercial software.

Finally, the proposed approach demonstrates that reliable substrate classification can be achieved using RGB imagery alone, offering a scalable alternative to multimodal workflows while adopting a more rigorous validation framework. For example, Rogers et al. (2025) reported high classification performance (F1 = 0.97) using combined lidar and RGB predictors, although independent validation reduced this to 0.81, highlighting the influence of validation strategy on reported accuracy. Similarly, Misiuk et al. (2021) demonstrated that non-spatial validation can substantially overestimate performance compared to spatially explicit approaches. This is relevant when comparing with RGB-based studies such as Arif et al. (2017), who reported accuracies of up to 97% for broad land-cover classes but only 61% for detailed grain-size classification, and Giroux et al. (2022), who achieved 79–86% accuracy for substrate delineation using object-based approaches. However, because these studies rely primarily on non-spatial validation strategies, their reported accuracies are likely influenced by spatial dependence between training and validation samples.

Within this context, the present framework achieves design-based Overall Accuracies ranging from 69.77% to 95.21% across sites, demonstrating comparable performance using RGB imagery alone under spatially explicit and independent validation. This is achieved by leveraging spectral- and texture-based features as proxies for surface structure, while ensuring that reported accuracies reflect true generalization performance. As a result, the method provides a scalable and robust alternative to multimodal approaches without requiring additional data sources.

5.4. Strategies for Improving Model Performance

The 0% User's Accuracy for `Dry Cobble (Large)` in the Sarca reach (Table 5) can be attributed to the extreme rarity of this class in the channelized morphology and the resulting sensitivity of the design-based estimator to small sample sizes. However, class frequency alone does not fully explain classification performance. For instance, `Dry Boulder`, despite being similarly underrepresented, achieved relatively high accuracy (81.82%), likely due to its more distinct spectral and structural characteristics, which reduce confusion with other substrate classes. In contrast, intermediate grain-size classes such as `Dry Cobble (Large)` are more difficult to distinguish because they occupy transitional positions within the continuous sediment size spectrum and share characteristics with adjacent classes.

 doi:10.20944/preprints202605.1410.v1

Overall, reduced classification accuracy for certain classes reflects a combination of class imbalance and the inherent continuity of sediment size transitions. Increasing the number and spatial diversity of training polygons for underrepresented classes would improve the model's ability to capture the full range of substrate variability. Where available, georeferenced field observations can also be used to guide the annotation process, helping to reduce interpreter uncertainty and potential annotator bias, particularly for visually ambiguous or transitional substrate classes.

Moreover, while Random Forest provided robust performance in this study, alternative machine learning algorithms may offer improvements depending on dataset characteristics. Rogers et al. (2025) evaluated multiple classifiers and found that performance varied depending on predictor structure and class separability. Ensemble and boosting methods designed to address class imbalance may further improve discrimination of minority substrate classes Seiffert et al. (2008).

Classification performance may also be influenced by the spatial scale at which texture features are computed. Texture descriptors summarize structural variability within a defined neighborhood, and the optimal scale depends on sediment organization. Smaller block sizes may better resolve fine-scale variability in homogeneous sediments, whereas larger blocks may improve discrimination in heterogeneous environments. Systematic evaluation of multiscale texture features represents an important direction for improving classification robustness and transferability.

5.5. Operational Implications and Limitations

The proposed workflow demonstrated high computational efficiency, enabling rapid classification of large UAV datasets and supporting frequent monitoring of river systems. For example, processing the Piave River dataset (368.17 ha) required approximately 22 minutes for annotation, 5 minutes for feature extraction, 60 minutes for full-domain classification, and 20 minutes for final validation. This scalability represents a significant advantage over grain-by-grain segmentation methods (e.g. Buscombe (2020); Mair et al. (2024)), which are computationally intensive and challenging to apply at reach scales.

Lower classification performance in the Brenta reach reflects the inherent difficulty of distinguishing substrate classes defined by gradual grain-size transitions. In this site, several classes occupy adjacent positions along the sediment-size continuum, increasing confusion between visually similar categories. These semantic boundaries are inherently subjective and may vary between annotators. Integration of SfM-derived roughness predictors and multi-operator annotation consistency analysis could improve classification reliability.

It is important to note that classification accuracy reflects agreement with expert interpretation rather than direct physical grain-size measurements. While this introduces semantic uncertainty, categorical classification provides a robust and operationally useful representation of substrate distribution. Unlike continuous grain-size estimation, categorical mapping is less sensitive to small measurement errors and enables consistent large-scale analysis.

The production of spatially continuous substrate maps enables reach-scale assessment of sediment organization, supporting applications such as habitat modeling, restoration planning, and geomorphic analysis. Similar applications have been demonstrated by Rogers et al. (2025), highlighting the broader utility of such mapping approaches beyond model evaluation.

Overall, these results demonstrate that ultra-high-resolution UAV imagery provides a physically meaningful and scalable basis for substrate classification. By combining spectral and structural predictors and implementing rigorous spatial validation, the proposed framework achieves reliable performance while significantly expanding accessibility compared to multimodal approaches.

6. Conclusion

This study presents a scalable and accessible framework for categorical substrate classification using ultra-high-resolution UAV-derived RGB imagery, integrating spectral information with statistical (GLCM) and structural (LBP) texture descriptors to enable reliable discrimination of substrate classes without relying on lidar or multimodal datasets. A key contribution is the explicit treatment of spatial

 doi:10.20944/preprints202605.1410.v1

autocorrelation (SAC) during model validation, demonstrating that standard cross-validation can substantially overestimate performance, while the combination of spatial cross-validation and design-based probability sampling provides both a conservative assessment of model generalization and an unbiased estimate of final map accuracy. Across four geomorphologically distinct river reaches, the method achieved design-based Overall Accuracies ranging from 69.77% to 95.21%, highlighting robust performance under varying conditions using RGB imagery alone. The workflow is implemented using open-source tools and standard GIS platforms, reducing barriers to adoption and enabling practical application for reach-scale mapping, with potential uses in habitat assessment, geomorphic analysis, and river management. Limitations include the requirement for clear water conditions and sensitivity to class imbalance and transitional sediment classes. Overall, the results demonstrate that accurate and operationally relevant substrate mapping can be achieved using RGB imagery alone when combined with appropriate feature representation and rigorous spatial validation, providing a practical alternative to more data-intensive approaches.

Disclaimer

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Exacutive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with the refinement and improvement of the English language and clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Acknowledgments: This research was funded by the Italian Ministry of Universities and Research (MUR) within the framework of the project DICAM-EXC – Department of Excellence 2023–2027 (grant L232/2016), and by the MUR PNRR project INEST – Interconnected Nord-Est Innovation Ecosystem (ECS00000043), funded by the NextGenerationEU programme. Additonal support was provided by the European Union's Horizon Europe Research and Innovation Programme through the Marie Skłodowska-Curie Postdoctoral Fellowship Programme (SMASH COFUND), grant agreement No. 101081355. The SMASH project is co-funded by the Republic of Slovenia and the European Union through the European Regional Development Fund.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

Arif, M. S. M., Gülch, E., Tuhtan, J. A., Thumser, P., & Haas, C. (2017). An investigation of image processing techniques for substrate classification based on dominant grain size using RGB images from UAV. *International Journal of Remote Sensing*, *38*(8-10), 2639–2661.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24–31.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Buscombe, D. (2013). Transferable wavelet method for grain-size distribution from images of sediment surfaces and thin sections, and other natural granular patterns. *Sedimentology*, *60*(7), 1709–1732.

Buscombe, D. (2020). SediNet: a configurable deep learning model for mixed qualitative and quantitative optical granulometry. *Earth Surface Processes and Landforms*, *45*(3), 638–651.

Carbonneau, P. E., Bergeron, N., & Lane, S. N. (2005). Automated grain size measurements from airborne remote sensing for long profile measurements of fluvial grain sizes. *Water resources research*, *41*(11).

 doi:10.20944/preprints202605.1410.v1

Carbonneau, P. E., Lane, S. N., & Bergeron, N. E. (2004a). Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery. *Water resources research*, *40*(7).

Carbonneau, P. E., Lane, S. N., & Bergeron, N. E. (2004b). Feature based image processing methods applied to bathymetric measurements from airborne remote sensing in fluvial environments. *Earth Surface Processes and Landforms*, *29*, 141–164. https://doi.org/10.1002/esp.1011.

Chardon, V., Schmitt, L., Piégay, H., & Lague, D. (2020). Use of terrestrial photosieving and airborne topographic LiDAR to assess bed grain size in large rivers: a study on the Rhine River. *Earth Surface Processes and Landforms*, *45*(10), 2314–2330.

Colditz, R. R. (2015). An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*, *7*(8), 9655–9681.

Farò, D., Soto Parra, T., Baumgartner, K., Andreoli, A., Vezza, P., & Zolezzi, G. (2025). An Integrated Framework for the Assessment of Meso-Scale Physical Habitats in Gravel-Bed Rivers Using Remote Sensing and 2D Hydraulic Modeling. *Wiley Interdisciplinary Reviews: Water*, *12*(3), e70027.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of environment*, *80*(1), 185–201.

Gevers, T., & Smeulders, A. W. (1999). Color-based object recognition. *Pattern recognition*, *32*(3), 453–464.

Giroux, C., Grant, J., Brown, C. J., & Barrell, J. (2022). Remote sensing of river habitat for salmon restoration. *Frontiers in Remote Sensing*, *3*, 993575.

Graham, D. J., Reid, I., & Rice, S. P. (2005). Automated sizing of coarse-grained sediments: image-processing procedures. *Mathematical geology*, *37*(1), 1–28.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, *67*(5), 786–804.

Knighton, D. (2014). *Fluvial forms and processes: A new perspective* (2nd ed.). Abingdon, Oxon; New York: Routledge. https://doi.org/10.4324/9780203784662.

Lang, N., Irniger, A., Rozniak, A., Hunziker, R., Wegner, J. D., & Schindler, K. (2020). GRAINet: mapping grain size distributions in river beds from UAV images with convolutional neural networks. *Hydrology and Earth System Sciences Discussions*, *2020*, 1–38.

Langhammer, J., & Vacková, T. (2018). Detection and mapping of the geomorphic effects of flooding using UAV photogrammetry. *Pure and Applied Geophysics*, *175*, 3223–3245. https://doi.org/10.1007/s00024-018-1874-1.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, *51*, 47–60.

Mair, D., Witz, G., Do Prado, A. H., Garefalakis, P., & Schlunegger, F. (2024). Automated detecting, segmenting and measuring of grains in images of fluvial sediments: The potential for large and precise data from specialist deep learning models and transfer learning. *Earth Surface Processes and Landforms*, *49*(3), 1099–1116.

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.

Misiuk, B., Lecours, V., Dolan, M., & Robert, K. (2021). Evaluating the suitability of multi-scale terrain attribute calculation approaches for seabed mapping applications. *Marine Geodesy*, *44*(4), 327–385.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971–987.

Parasiewicz, P. (2007, 10). The MesoHABSIM model revisited. *River Research and Applications*, *23*, 893-903. https://doi.org/10.1002/rra.1045.

Piégay, H., Arnaud, F., Belletti, B., Bertrand, M., Bizzi, S., Carbonneau, P., Dufour, S., Liébault, F., Ruiz-Villanueva, V., & Slater, L. (2020). Remotely sensed rivers in the Anthropocene: State of the art and prospects. *Earth Surface Processes and Landforms*, *45*(1), 157–188.

Pietikäinen, M., & Zhao, G. (2015). Two decades of local binary patterns: A survey. In *Advances in independent component analysis and learning machines* (pp. 175–210). Elsevier.

Powell, D. M. (1998). Patterns and processes of sediment sorting in gravel-bed rivers. *Progress in physical geography*, *22*(1), 1–32.

Purinton, B., & Bookhagen, B. (2019). Introducing PebbleCounts: a grain-sizing tool for photo surveys of dynamic gravel-bed rivers. *Earth Surface Dynamics*, *7*(3), 859–877.

 doi:10.20944/preprints202605.1410.v1

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Rogers, J. M., Brasington, J., & Hoyle, J. (2025). Bed material facies mapping at braided river scale and evidence for trends in fine sediment. *Earth Surface Processes and Landforms*, *50*(2), e70012.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. In *2008 19th international conference on pattern recognition* (pp. 1–4).

Stehman, S. V. (2012). Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sensing Letters*, *3*(2), 111–120.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 25.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(sup1), 234–240.

Verdú, J. M., Batalla, R. J., & Martínez-Casasnovas, J. A. (2005). High-resolution grain-size characterisation of gravel bars using imagery analysis and geo-statistics. *Geomorphology*, *72*(1-4), 73–93.

Vezza, P., Parasiewicz, P., Spairani, M., & Comoglio, C. (2014). Habitat modeling in high-gradient streams: the mesoscale approach and application. *Ecological Applications*, *24*, 844-861.

Wadoux, A. M.-C., Brus, D. J., & Heuvelink, G. B. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, *355*, 113913.

Wadoux, A. M.-C., Heuvelink, G. B., De Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692.

Wolman, M. G. (1954). A method of sampling coarse river-bed material. *Transactions, American Geophysical Union*, *35*(6), 951–956. https://doi.org/10.1029/TR035i006p00951.

Woodget, A. S., & Austrums, R. (2017). Subaerial gravel size measurement using topographic data derived from a UAV-SfM approach. *Earth Surface Processes and Landforms*, *42*(9), 1434–1443.

Woodget, A.S., Visser, F., Maddock, I. P. E. Carbonneau, P., Austrums, R. (2016, February 7-12). Quantifying fluvial substrate size using hyperspatial resolution UAS imagery and SfM-photogrammetry. In *Proceedings of the 11th international symposium on ecohydraulics.* Melbourne, Australia. (Extended Abstract)