

Computational Discovery and Intelligent Systems CDIS

ISSN: 3070-5037/© 2026, The Authors. CDIS (CC BY 4.0) License.

Journal Homepage

<https://pub.scientificirg.com/index.php/CDIS>



Reproducible Spectral-Spatial Vegetation Stratification from a Single Jilin-1 Scene under Controlled Validation

Thanaa Moustafa^a , Mohammed Melhi^b, and A. A. Somaie^{c,1}

^a Department of Information System, Suez Canal University, Ismaila, Egypt; E-mail: sanaamostafa59@gmail.com

^b PhD, University of Bradford, UK, and Chief Executive Officer CEO of Rushd AI Company and Amatrix AI Company, Riyadh, KSA;

E-mail: mmelhi@rushdai.com, mmelhi@promatrix.ai

^c PhD, University of Bradford, UK, PDF Post-Doctoral Fellow Research Associate, University of Calgary, Canada, Software Engineering Program SE, Faculty of Computer Science, October University for Modern Sciences & Arts MSA, 6 October, Giza, Egypt; E-mail: alibrahim@msa.edu.eg, aasomaie@gmail.com

ABSTRACT

Very-high-resolution (VHR) satellite imagery may support individual palm delineation and field-calibrated assessment of vegetation condition, but such programs require costly multi-farm imagery, tree-level annotations, and field measurements. This study asks whether a single Jilin-1 scene, acquired on 6 October 2024, contains enough reproducible spectral-spatial structure to justify that investment. The scene serves only as a technical test bed because palm occurrence within it was not verified. An exploratory Stage 1 pipeline (Brovey preprocessing, four spectral indices, three-cluster MiniBatch K-Means pseudo-labels, and a U-Net) reached 0.9806 agreement with its own labels, but redundant features, whole-scene clustering, overlap between clustering features and model inputs, and reuse of validation data limited interpretation. A controlled Stage 2 removed an exact duplicate (green-NIR normalized contrast) and an algebraic duplicate (ratio vegetation index), clustered on NDVI and GNDVI, and selected two clusters from a five-seed analysis (silhouette 0.59472; pairwise adjusted Rand index 0.99034). Training, validation, and held-out test regions were separated by guard patches, and the scaler and clustering model were fitted on training data only. In the held-out region, a bands-only Logistic Regression reproduced the two-stratum reference with 0.99401 agreement, compared with 0.99192 ± 0.00153 for a bands-only U-Net over three seeds. The U-Net did not exceed Logistic Regression in agreement, although its macro-F1 and mean intersection over union were higher in point estimate. Supplying NDVI and GNDVI to the U-Net improved every metric, which provides direct evidence of target-feature overlap. The scene therefore contains strong, reproducible, scene-relative spectral structure that raw bands recover in a geographically separated region, and a deep network is not required for high pseudo-label reproduction. Because Brovey fusion degraded reduced-resolution radiometric fidelity, these results are conditional on that preprocessing choice. The strata are not validated indicators of moisture, stress, disease, or palm condition, and a multi-farm, multi-date study with palm annotations and field measurements remains necessary.

PAPER INFORMATION

HISTORY

Received: 18 August 2026

Revised: 1 October 2026

Accepted: 16 October 2026

Online: 31 October 2026

MSC

68U10; 68T10; 62H30; 68T07

KEYWORDS

Spectral-Spatial;
Vegetation Mapping;
Jilin-1 Scene;
Weak Supervision;
Palm Monitoring.

¹Corresponding author at PhD, University of Bradford, UK, PDF Post-Doctoral Fellow Research Associate, University of Calgary, Canada, Software Engineering Program SE, Faculty of Computer Science, October University for Modern Sciences & Arts MSA, 6 October, Giza, Egypt; E-mail: alibrahim@msa.edu.eg, aasomaie@gmail.com

1 INTRODUCTION

Large-area vegetation monitoring comprises several related but distinct tasks. Vegetation detection separates vegetated from non-vegetated surfaces, and vegetation-condition mapping describes relative spectral or structural differences within vegetation. Object delineation divides continuous canopy into individual crowns or regions, and disease diagnosis assigns a biologically verified cause. Each task requires a stronger form of reference information than the one before it, so the tasks are not interchangeable.

Field surveys provide biologically meaningful observations but are laborious and difficult to scale. Satellite imagery complements them by recording visible and near-infrared response over large areas. Pigment content, canopy density, leaf structure, illumination, soil background, and water status all modify this response, and similar spectra can arise from several biotic and abiotic causes. A spectral anomaly in satellite data therefore cannot be read as evidence of a particular disease, pest, or water deficit.

The long-term motivation is to determine whether VHR satellite imagery can support individual-tree palm monitoring, including crown delineation and field-calibrated assessment of vegetation condition. Such a system would have to delineate individual palms, calibrate moisture, water status, or stress against field measurements, and transfer to additional farms and acquisition dates. Published palm studies illustrate the reference data that these steps require: annotated crown mapping from VHR satellite data [1], manually annotated palm-health segmentation from UAV imagery [2, 3], and field-observed Red Palm Weevil monitoring [4]. Assembling comparable references across several farms and dates is expensive, so a prior question deserves an answer. Does the available satellite imagery contain enough reproducible spectral-spatial structure to make such a campaign technically worthwhile? The present study evaluates only this initial technical prerequisite.

No palm-specific claim is evaluated in this study. The analyzed scene is used solely as a technical test bed for spectral-spatial learnability and validation design. Because palm occurrence within the scene was not verified, the study does not evaluate palm detection, crown delineation, or palm-condition assessment, and references to palm monitoring describe the intended future application, not a demonstrated property of the study area. The image-derived strata reported below are scene-relative spectral groupings. None of them is a validated indicator of moisture, water status, stress, disease, or palm condition, and this scope statement applies to every result in the paper. Because field labels were unavailable, the feasibility question is approached through weak, image-derived supervision. A clustering model defines the strata, and classifiers and segmentation networks are scored by how well each reproduces the strata in a region that did not contribute to defining them.

Two completed stages are reported. Stage 1 is an exploratory proof of concept that combined Brovey preprocessing, four spectral indices, three-cluster MiniBatch K-Means, and a U-Net, and it produced high same-scene agreement. Its interpretation is limited by redundant features, an unsupported cluster number, clustering fitted on the whole scene before data splitting, overlap between clustering features and network inputs, spatial dependence between training and validation patches, reuse of the validation subset for evaluation, a single training run, absent input normalization, and the lack of simple baselines. Stage 2 is a controlled experiment designed to test whether the exploratory result survives once these dependencies are controlled. Stage 2 provides the primary evidence of this paper, and Stage 1 is reported concisely as a transparent control example.

The contributions are as follows:

- (i) a transparent account of the Stage 1 pipeline and of the dependencies that limit its interpretation;
- (ii) a controlled Stage 2 experiment that combines a training-region redundancy audit, repeated-seed cluster-number selection, a guarded spatial partition, train-only clustering and normalization, simple and deep model comparisons, input ablations, and spatial-block bootstrap uncertainty; and
- (iii) an explicit assessment of what the resulting evidence does and does not establish for a field-validated individual-palm monitoring study.

The study addresses five research questions:

- RQ1: Does the Jilin-1 scene contain reproducible scene-relative spectral structure after redundant features are removed?
- RQ2: Which cluster configuration is best supported in the training region under repeated-seed sensitivity analysis?
- RQ3: Can spectral strata generated from the training region be reproduced in a geographically separated held-out region from raw spectral bands alone?
- RQ4: Does a spatial deep-learning model provide a measurable benefit over simple non-spatial spectral classification?
- RQ5: Do the observations provide sufficient technical evidence to motivate a multi-farm, field-validated individual-palm monitoring study?

RQ5 concerns technical feasibility only, and answering it does not establish that any stratum corresponds to a biological or physiological condition.

The remainder of the paper is organized as follows. Section 2 reviews palm-monitoring studies, weak supervision, spatial validation, and image fusion. Section 3 describes the data, preprocessing, the Stage 1 baseline, and the controlled Stage 2 design. Section 4 reports the redundancy audit, cluster-number selection, model comparison, bootstrap uncertainty, and the fusion assessment. Section 5 interprets the findings against the research questions, Section 6 states the limitations and future work, and Section 7 concludes.

2 RELATED WORK

Vegetation indices condense combinations of spectral bands into descriptors of relative canopy response. NDVI contrasts red absorption with near-infrared reflectance [5], the ratio vegetation index (RVI) expresses the near-infrared-to-red ratio [6], and GNDVI replaces the red band with the green band [7]. These indices are indirect descriptors, and their interpretation depends on vegetation type, season, illumination, atmospheric conditions, soil background, and sensor preprocessing. The green and near-infrared normalized difference with the opposite sign is historically known as NDWI and was proposed for open-water delineation [8], not for measuring vegetation water content. Because several indices share the NDWI acronym, this study uses the descriptive term *Green-NIR Normalized Contrast* (GNC) for the corresponding Stage 1 feature, and GNC is not interpreted as a measure of vegetation water content.

Independent biological labels become necessary once the objective changes from spectral stratification to biological classification. **Table 1** summarizes the studies most relevant to palm monitoring and to the methodological issues considered here, together with the limitation of each study relative to the present objective. Date-palm crowns have been segmented from WorldView-3 imagery with transformer architectures, and multispectral input improved mean intersection over union (mIoU) relative to RGB input [1]. Multiscale vision transformers applied to UAV imagery have supported palm instance segmentation and a healthy-versus-unhealthy assessment [2]. A two-year UAV experiment with RGB, multispectral, and thermal sensors found that thermal statistics distinguished severe Red Palm Weevil infestation before visible canopy symptoms [4]. Manually annotated UAV tiles have also been used to segment healthy palm, unhealthy palm, and background in a Moroccan oasis [3]. These studies rely on independent annotations or field observations that the present scene does not provide. The studies indicate the reference information required for later biological monitoring, but offer no evidence about vegetation identity or condition in the study area.

Table 1: Summary of related studies and their limitations relative to the present objective

Reference	Year	Data	Task and method	Key result	Limitation
[1]	2024	WorldView-3 satellite imagery of date palms	Crown segmentation with several transformer-based segmentation architectures	Multispectral input improved mIoU over RGB; UperNet-Swin reached 78.10% mIoU	Relies on manually delineated palm references and commercial VHR imagery; addresses crown mapping and counting, not field-calibrated moisture or stress
[2]	2024	UAV imagery of date-palm plantations	Multiscale vision transformer with Mask R-CNN for detection, segmentation, and condition classification	F-scores of 94.33% (detection), 94.2% (segmentation), and 88.4% (condition)	UAV acquisition limits comparability with satellite-scale monitoring; condition labels rest on visual annotation, not field measurements of plant water status
[4]	2023	Two years of UAV RGB, multispectral, and thermal imagery of an experimental field	Red Palm Weevil detection with segmentation and vegetation-index analysis	Thermal statistics separated severely infested palms before visible symptoms	Single experimental field, and the authors state that further validation is needed; no thermal band exists in the Jilin-1 scene
[3]	2026	UAV RGB tiles from the Figuig oasis (296 annotated tiles)	U-Net and DeepLabV3+ for healthy palm, unhealthy palm, and background	Macro-F1 of 82.06% (DeepLabV3+) and 81.28% (U-Net)	The heterogeneous unhealthy class was hard to segment; limited dataset diversity and RGB-only information
[9]	2024	ISPRS Potsdam, ISPRS Vaihingen, and iSAID	Weak supervision with a self-supervised encoder and a noise-suppressing loss	Performance close to fully supervised methods	Pseudo-label quality remains the central dependency; targets generic land-cover classes, not calibrated agricultural condition
[10]	2024	DLRS, Potsdam, and an oil-field dataset	Text-guided weak supervision with fine-tuned CLIP, background suppression, and SAM edge refinement	Outperformed other weakly supervised methods	Depends on pretrained vision-language representations; evaluated on object categories, not field-validated crop condition
[11]	2025	Remote-sensing images	Semi-supervised segmentation with enhanced guidance and pseudo-label optimization	Addresses error accumulation from inaccurate pseudo-labels	Still relies on image-derived pseudo-labels and benchmark classes; no link to independent biological measurements

Table 1 (continued)

Reference	Year	Data	Task and method	Key result	Limitation
[12]	2017	Spatially, temporally, or hierarchically structured data	Review of cross-validation strategies for structured data	Blocked designs account for dependence	General guidance from ecological applications; no partition rule for dense image segmentation
[13]	2019	Species distribution data	blockCV package for spatially or environmentally separated folds	Software support for blocked validation	Built for species distribution models, not for dense image patches
[14]	2021	Multispectral and panchromatic images from several sensors	Benchmark of classical and emerging pansharpening methods	Common protocols for method comparison	Focuses on fusion quality; effects on downstream classification or segmentation are not examined
[15]	2024	WorldView-3 imagery	Unsupervised CNN pansharpening with balanced spectral and spatial losses	Stable training through loss balancing	Assessed with visual quality and numerical indexes; downstream effects are not examined
[16]	2025	QuickBird, GaoFen-2, and WorldView-3 imagery	Hybrid transformer pansharpening (HSSTN)	Sharper spatial detail with fewer spectral artifacts	Assessed with quantitative metrics and visual quality; downstream effects are not examined

Semantic segmentation in remote sensing commonly relies on encoder-decoder networks. U-Net introduced skip connections that preserve spatial detail [17], and EfficientNet provides scalable convolutional encoders [18]. Deep learning has since been applied across a broad range of remote-sensing interpretation tasks [19]. Large-scale pretraining on spectral imagery, including hyperspectral foundation models [20], represents a further direction, although all networks in this study are trained from scratch.

Segmenting image-derived pseudo-labels differs from segmenting against independent annotations. High agreement with a clustering reference shows that a model is consistent with that reference, but it does not establish biological validity or transferability beyond the scene. Weak supervision replaces or supplements manual labels with labels obtained from heuristics, clustering, prior models, or other indirect sources. Recent remote-sensing methods address pseudo-label quality through self-supervised encoders, noise-robust losses, text-guided activation maps, and consistency regularization [9, 10, 11]. A principal risk arises when the supervision source shares information with the model inputs. If the same spectral indices define the pseudo-labels and enter the prediction model, agreement can be inflated because the model has direct access to the variables used to generate its targets.

Spatial dependence raises a further validation concern. Non-overlapping patches from the same image can remain environmentally and visually similar, and a clustering model fitted on the complete scene links training and evaluation regions. Spatially structured data therefore require validation designs that separate regions or blocks in space [12, 13]. Pansharpening adds an interpretive constraint because spectral and spatial fidelity can trade off, and a finer sampling grid does not necessarily provide independent fine-scale multispectral information [14, 15, 16].

Taken together, the literature supplies palm-monitoring studies based on independent labels, methods for weak supervision, and guidance on spatial validation and image fusion. It offers little evidence on whether a single VHR scene without field labels contains enough learnable structure to justify further data collection. The present study addresses this gap through a controlled design that follows the validation principles outlined above.

3 MATERIALS AND METHODS

3.1 Study area and input data

The study scene is located in southern New Zealand and contains heterogeneous vegetation. The geodesic area of the supplied WGS84 study-boundary polygon is 27.2355 km². The 109,216,876 nonzero panchromatic pixels, at 0.25 m² each (0.5 m×0.5 m), cover 27.304 km², which agrees with the polygon area to within 0.3%. The rectangular 16026 × 11833 bounding box includes no-data margins and is not a valid study-area measurement.

The imagery was acquired by the JL1KF01C satellite with the PMSL1 line push-broom sensor on 6 October 2024, starting at 06:37:51. Recorded center times were 06:37:52 for the multispectral product and 06:37:53 for the panchromatic product, and both products were distributed at processing level L1. Product metadata reported approximately 1% cloud cover, a solar azimuth of approximately 321.19°, a solar elevation of approximately 42.73°, and a satellite elevation of approximately 78.4°. The panchromatic (PAN) raster contained 16,026 × 11,833 pixels in one band with metadata-reported row and column ground sampling distances (GSDs) of 0.5 m. The four-band multispectral (MSS) raster contained 4,008 × 2,959 pixels with metadata-reported GSDs of 2.0 m in the row direction and 2.00455 m in the column direction. Both rasters were stored as unsigned 16-bit integers. Among non-black native MSS pixels, the raw digital number (DN) ranges were 132 to 1171 for blue, 129 to 1732 for green, 46 to 1912 for red, and 158 to 3014 for NIR. The 0.5 m value refers to the

PAN sampling interval, whereas the multispectral information is sampled at approximately 2 m. **Table 2** lists the product metadata and raster properties.

Table 2: Jilin-1 product metadata and raster properties

Parameter	Multispectral	Panchromatic
Satellite and sensor	JL1KF01C, PMSL1	JL1KF01C, PMSL1
Product level	L1	L1
Acquisition date	2024-10-06	2024-10-06
Start time	06:37:51	06:37:51
Center time	06:37:52	06:37:53
Dimensions (pixels)	4008 × 2959	16026 × 11833
Bands	4	1
Band spectral ranges	Not documented	Not documented
Storage and unit	uint16, uncalibrated DN	uint16, uncalibrated DN
Row GSD (metadata)	2.0 m	0.5 m
Column GSD (metadata)	2.004 55 m	0.5 m
Cloud cover	1%	1%
Solar azimuth	321.188°	321.189°
Solar elevation	42.7266°	42.7267°
Embedded RPC	Yes	Yes
Affine CRS read by Rasterio	None, identity-like transform	None, identity-like transform
Package projection files	WGS84 definition	WGS84 definition

An explicit no-data audit, presented in **Table 3**, showed that 42.4070% of PAN bounding-box pixels and 42.3618% of native MSS pixels were black margin. Proportions computed over the full bounding box cannot describe physical scene composition, so all class proportions reported below refer to retained patches within explicitly defined partitions.

Table 3: Valid and black-margin pixels in the raster bounding boxes

Raster condition	Pixels	Total pixels	Percentage
PAN: PAN > 0	109,216,876	189,635,658	57.5930%
PAN: PAN = 0	80,418,782	189,635,658	42.4070%
MSS: any band > 0	6,835,696	11,859,672	57.6382%
MSS: all bands = 0	5,023,976	11,859,672	42.3618%

The product package included WGS84 projection-definition files and embedded rational polynomial coefficient (RPC) metadata for both PAN and multispectral imagery. The TIFF rasters used in the analysis, however, did not expose a map-projected coordinate reference system or a valid affine transform through Rasterio and were read with CRS=None and an identity-like transform. Both stages therefore operated in image coordinates. The RPC model was not used for orthorectification, and PAN and MSS co-registration error was not measured, so the physical footprint of an image patch is not asserted.

The Level-1 metadata also specified the vendor radiance relation of **Equation 1**,

$$L = \text{DN} \cdot G_{\text{abs}} + B_{\text{abs}}, \quad (1)$$

where DN is the raw digital number, and G_{abs} and B_{abs} are band-specific absolute gain and bias coefficients and L is spectral radiance in $\text{W m}^{-2} \text{sr}^{-1} \mu\text{m}^{-1}$. These coefficients were available but were not applied in either stage. All processing used raw DN-domain values without conversion to radiance or surface reflectance and without atmospheric correction, so the resulting strata are scene-relative and are not radiometrically standardized measurements that transfer across dates or sensors.

3.2 Overall study design

Figure 1 shows the three-stage structure of the work. Stage 1 is the exploratory proof of concept, and Stage 2 is the controlled experiment that provides the primary evidence. Stage 3 lists validation and deployment steps that were not performed and are required in future work. Both completed stages share the data, preprocessing, feature definitions, and patch grid described in Sections 3.3 and 3.4.

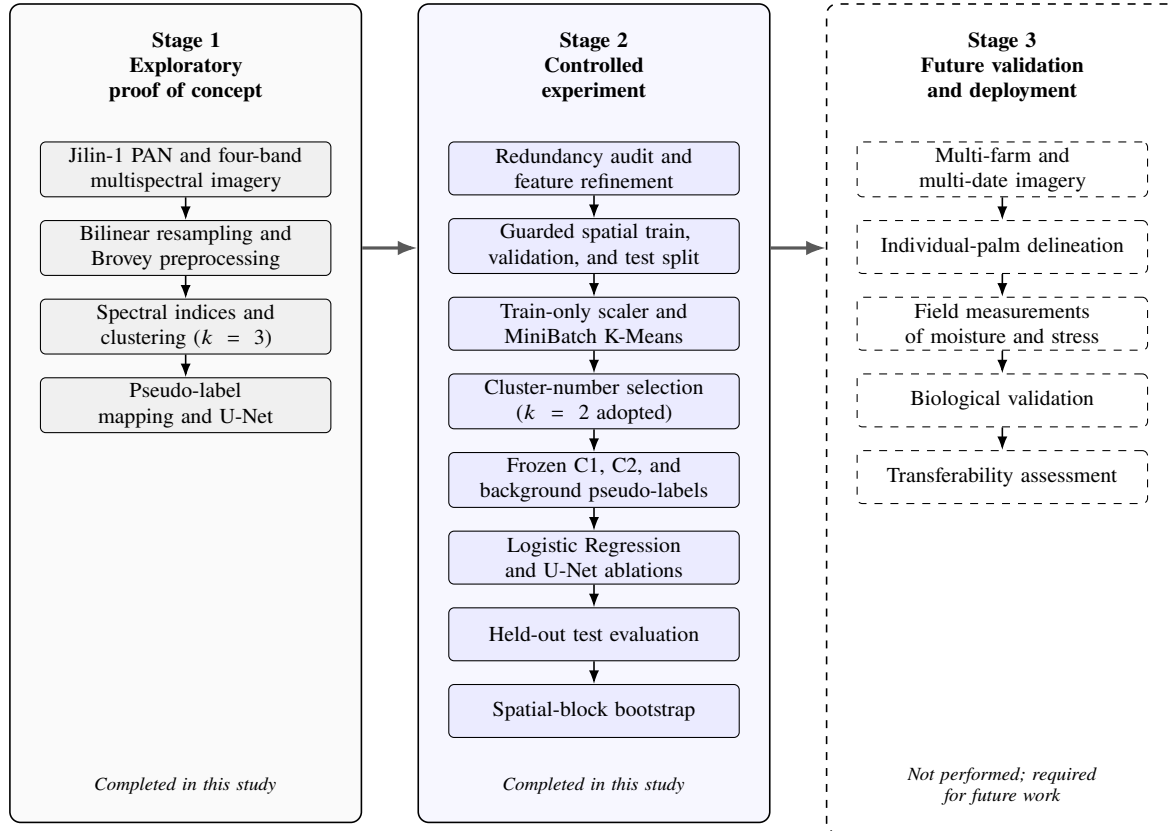


Figure 1: Three-stage study design

Table 4 pairs each methodological dependency identified in Stage 1 with the control introduced in Stage 2.

Table 4: Stage 1 dependencies and the corresponding Stage 2 controls

Stage 1 dependency	Stage 2 control
Feature redundancy (GNC and GNDVI; RVI and NDVI)	Training-region redundancy audit; GNC and RVI removed; clustering on NDVI and GNDVI
Unsupported cluster number ($k = 3$)	Five-seed silhouette and pairwise ARI analysis for $k = 2, \dots, 6$; $k = 2$ adopted
Clustering fitted on a whole-scene sample	Scaler and MiniBatch K-Means fitted on training data only, then frozen
Target-feature overlap	Bands-only models as primary comparisons; indices-only and bands-plus-indices U-Nets as overlap ablations
Spatial dependence between subsets	Spatially separated regions with guard patches; spatial-block bootstrap
Validation reused for evaluation	Validation region for checkpoint selection only; separate held-out test region
Single training run	Three fixed seeds per network configuration; five seeds for clustering analysis
No input normalization	Per-channel normalization computed from training patches only
No simple baseline	Bands-only Logistic Regression; 3×3 smoothing diagnostic
Untested Brovey fidelity	Reduced-resolution SAM and ERGAS comparison with a no-fusion baseline

3.3 Resampling and Brovey preprocessing

Both stages first resampled the MSS bands to the PAN grid with bilinear interpolation, leaving the PAN raster at its native resolution, and then applied Brovey fusion to the DN-domain values. Brovey fusion was retained to preserve comparability with the exploratory Stage 1 pipeline and to isolate the effect of the methodological controls from changes

in preprocessing, not because a comparative study had shown it to be the best choice. For multispectral band M_i , the implemented transformation was

$$M'_i = \frac{M_i}{\sum_{j=1}^4 M_j + \epsilon} P, \quad (2)$$

where P is the PAN value and $\epsilon = 10^{-10}$ avoids division by zero. Ratio-based indices were computed after fusion. Defining the pixel-wise factor of **Equation 3**,

$$c = \frac{P}{\sum_{j=1}^4 M_j + \epsilon}, \quad (3)$$

Equation 2 gives $M'_i = cM_i$ for every band, so that for NDVI

$$\frac{\text{NIR}' - R'}{\text{NIR}' + R'} = \frac{c \text{NIR} - c R}{c \text{NIR} + c R} = \frac{\text{NIR} - R}{\text{NIR} + R}, \quad (4)$$

apart from the numerical stabilizers. The same cancellation holds for GNDVI and for the ratio RVI. Brovey fusion therefore enlarged the raster grid without creating independent sub-meter multispectral measurements, and indices computed from fused bands carry the same information as indices computed from the resampled MSS bands. Only the vegetation mask of Section 3.4, which thresholds the fused NIR band, depends on the PAN intensity. **Figure 2** compares native and Brovey-processed RGB composites and NDVI maps.

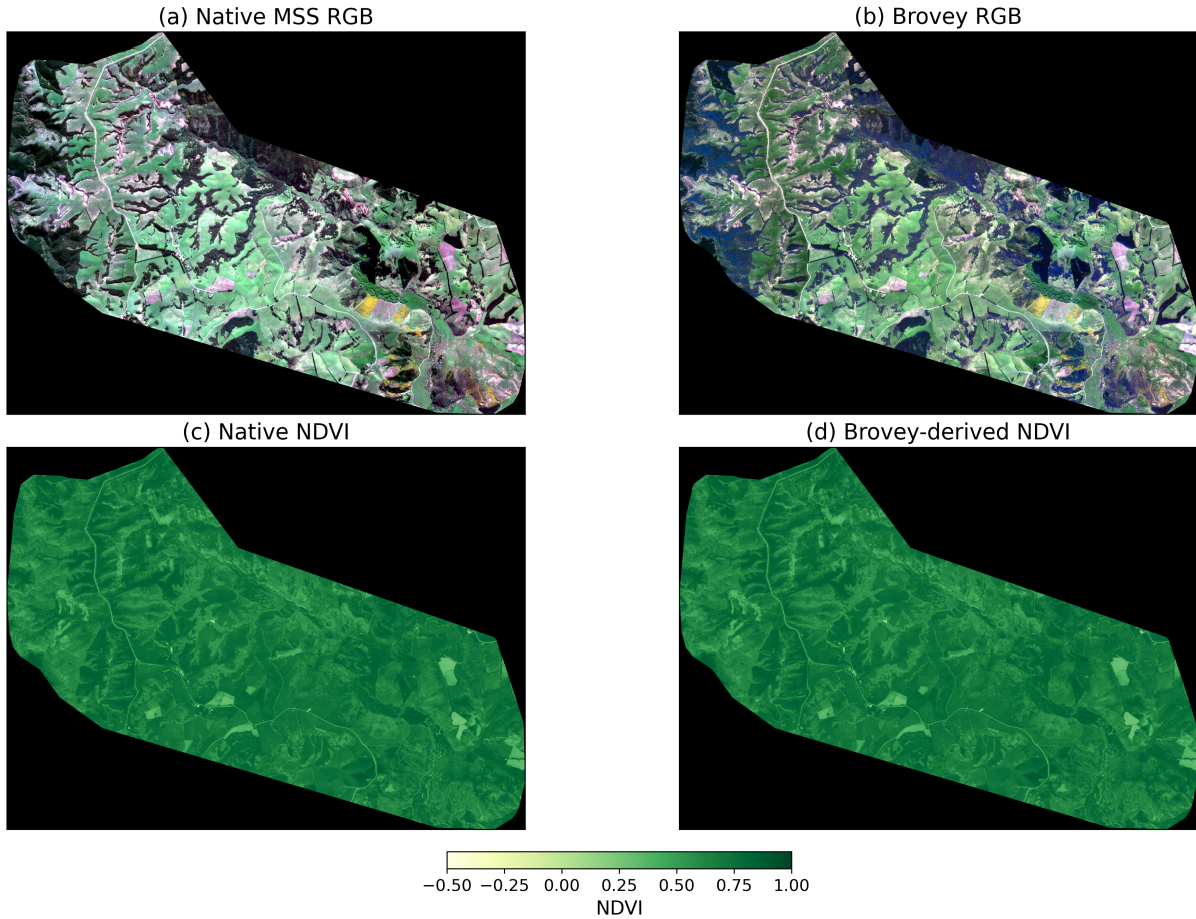


Figure 2: Native and Brovey-processed RGB composites and NDVI maps

In **Figure 2**, panels (a) and (b) show the native MSS and Brovey RGB composites, and panels (c) and (d) show the corresponding NDVI maps on one shared color scale. Black areas are no-data margins outside the valid footprint. The Brovey composite is visibly darker and bluer in shaded terrain than the native composite, which is consistent with the rescaling of each band by the PAN-to-MSS-sum ratio. The two NDVI maps are visually indistinguishable at the displayed scale, as **Equation 4** predicts, and most of the scene falls at moderately high NDVI. Both maps are computed from uncalibrated DN values and show relative spectral variation only.

A reduced-resolution assessment compared the bilinear no-fusion baseline with Brovey fusion over 6,835,681 valid evaluation pixels, a count close to the number of nonzero MSS pixels in **Table 3**. The protocol follows standard pansharpening

evaluation [14, 15], with the original multispectral image as reference. Spectral fidelity was summarized by the spectral angle mapper (SAM, **Equation 5**) [21, 14],

$$\text{SAM} = \arccos\left(\frac{\langle \mathbf{x}, \hat{\mathbf{x}} \rangle}{\|\mathbf{x}\| \|\hat{\mathbf{x}}\|}\right), \quad (5)$$

averaged over valid pixels, where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are the reference and fused band vectors, and by the relative dimensionless global error in synthesis (ERGAS, **Equation 6**) [22, 14],

$$\text{ERGAS} = 100 \frac{h}{l} \sqrt{\frac{1}{N_b} \sum_{b=1}^{N_b} \left(\frac{\text{RMSE}_b}{\mu_b} \right)^2}, \quad (6)$$

where h/l is the ratio of PAN to MSS pixel sizes (0.25 for the 0.5 m and 2 m sampling), N_b is the number of bands, RMSE_b is the root-mean-square error of band b , and μ_b is the reference mean of band b . Lower values of both indicators denote better fidelity.

3.4 Spectral features, vegetation mask, and patch extraction

Four candidate derived features were computed from the Brovey bands: NDVI [5], GNC (the green and NIR normalized difference of [8]), GNDVI [7], and RVI [6], defined as

$$\text{NDVI} = \frac{\text{NIR} - R}{\text{NIR} + R + \epsilon}, \quad (7)$$

$$\text{GNC} = \frac{G - \text{NIR}}{G + \text{NIR} + \epsilon}, \quad (8)$$

$$\text{GNDVI} = \frac{\text{NIR} - G}{\text{NIR} + G + \epsilon}, \quad (9)$$

$$\text{RVI} = \frac{\text{NIR}}{R + \epsilon}, \quad (10)$$

where $\epsilon = 10^{-10}$ and RVI was not clipped. Ignoring the stabilizer, **Equation 8** reduces to

$$\text{GNC} = \frac{G - \text{NIR}}{G + \text{NIR}} = -\frac{\text{NIR} - G}{\text{NIR} + G} = -\text{GNDVI}, \quad (11)$$

and an empirical check confirmed exact duplication in the implemented arrays (maximum and mean $|\text{GNC} + \text{GNDVI}|$ both 0, and $\text{corr}(\text{GNC}, \text{GNDVI}) = -1.0$). With $\text{RVI} = \text{NIR}/R$,

$$\text{NDVI} = f(\text{RVI}) = \frac{\text{RVI} - 1}{\text{RVI} + 1}, \quad (12)$$

so RVI is a monotone algebraic transform of NDVI apart from the stabilizers. **Table 5** summarizes the use of each feature in the two stages.

Table 5: Spectral features and their use in the two stages

Feature	Interpretation	Stage 1 use	Stage 2 use
B, G, R, NIR	Brovey-processed band responses (DN)	U-Net input	Input to bands-only and bands-plus-indices models
NDVI	Normalized contrast of NIR and red	Clustering and U-Net input	Clustering; input to indices-only and bands-plus-indices U-Nets
GNDVI	Normalized contrast of NIR and green	Clustering and U-Net input	Clustering; input to indices-only and bands-plus-indices U-Nets
GNC	Exact negative of GNDVI	Clustering and U-Net input	Removed (exact duplicate)
RVI	Unclipped NIR-to-red ratio; algebraic transform of NDVI	Clustering and U-Net input	Removed (algebraic duplicate)

The operational image-domain vegetation mask of both stages was

$$m(p) = \mathbf{1}[\text{NDVI}(p) > 0.15 \wedge \text{NIR}'(p) > 50 \wedge \text{all features finite}], \quad (13)$$

and pixels with $m(p) = 0$ formed the background class. The mask is a processing rule and not biological ground truth, and the value 50 is a processed-image threshold, not a sensor-calibrated physiological threshold. In a large random non-black

sample, retention after the NDVI condition changed by only about 0.13 percentage points between processed-NIR thresholds of 25 (99.9872%) and 50 (99.8523%), and it fell to 88.4066% at 100. This is a robustness observation and not a justification of the value 50.

The processed image stack and the label raster were divided into non-overlapping 256×256 -pixel patches (stride 256) in row-major order. Patches with less than 10% non-background pixels were discarded, which left 1,723 retained patches, and Stage 2 reproduced this patch set exactly.

3.5 Stage 1: exploratory proof of concept

Stage 1 clustered pixels that passed the vegetation mask using the feature vector $Z_1 = [\text{NDVI}, \text{GNC}, \text{GNDVI}, \text{RVI}]$. Up to 500,000 valid vegetation pixels were sampled from the complete scene without replacement after setting the NumPy seed to 42. A `StandardScaler` was fitted to this sample, and `MiniBatch K-Means` [23] was fitted with $k = 3$, `random_state=42`, `batch_size=4096`, and `n_init=10`, with other settings at the library defaults of an unrecorded scikit-learn version [24]. The value $k = 3$ was an operational choice for a coarse three-level stratification, not a statistically optimized cluster number. The clusters were ordered by mean NDVI into higher-, intermediate-, and lower-NDVI strata, which together with background gave four pseudo-label classes. The labels C1 and C2 are reserved for the Stage 2 strata.

The Stage 1 network was a U-Net [17] implemented with `segmentation_models_pytorch` and an EfficientNet-B4 encoder [18] without pretrained weights. It received the eight-channel stack $X_1 = [B, G, R, \text{NIR}, \text{NDVI}, \text{GNC}, \text{GNDVI}, \text{RVI}]$ without per-channel normalization and predicted four classes. The 1,723 retained patches were split in row-major order into 1,378 training and 345 validation patches. Training used batch size 8, 30 epochs, AdamW [25] with learning rate 3×10^{-4} and weight decay 1×10^{-4} , cosine annealing, and random horizontal flips, vertical flips, and 90° rotations. The loss was the weighted cross-entropy of **Equation 14**,

$$\mathcal{L} = - \frac{\sum_p w_{y_p} \log \hat{\pi}_{p,y_p}}{\sum_p w_{y_p}}, \quad (14)$$

where y_p is the label of pixel p , $\hat{\pi}_{p,c}$ is the predicted probability of class c at pixel p , and w_c is the weight of class c . Stage 1 used the heuristic weights $[1.0, 1.0, 1.0, 0.2]$, which down-weight the background class. A checkpoint was saved whenever the validation loss decreased, and the saved checkpoint was later evaluated on the same 345 validation patches.

Four features of this design limit the interpretation of Stage 1. The clustering model was fitted on a whole-scene sample before the split, so validation pixels influenced the pseudo-label definition. The clustering indices were also network inputs, the same patches served for checkpoint selection and evaluation, and only one training run was available. Stage 1 results are therefore reported in Section 4.1 for transparency only.

3.6 Stage 2: controlled experiment

Algorithm 1 summarizes the Stage 2 pipeline, and the following subsections give the details of each step.

3.6.1 Feature redundancy audit

Redundancy among the candidate features MS Blue, MS Green, MS Red, MS NIR, NDVI, GNDVI, and RVI was audited on a sample from the training region only. GNC was excluded from the audit and from all Stage 2 analyses because it is the exact negative of GNDVI, as shown in **Equation 11**. Pairwise Spearman rank correlations were computed, and the suspected relation between NDVI and RVI was tested directly through the maximum and mean absolute deviation between NDVI and $f(\text{RVI})$ of **Equation 12**. A feature was removed only when it was an exact or algebraic duplicate of another feature, and high correlation alone was not treated as grounds for removal.

3.6.2 Guarded spatial partition

Before any scaler, clustering model, or normalization statistic was computed, the 1,723 retained patches were assigned to spatially separated training, validation, and held-out test regions in image coordinates, as shown in **Figure 3**. The partition was made along the image column (x) axis, with the training region on the left, the validation region in the center, and the test region on the right. Guard patches were excluded on both sides of each boundary, which gave a guard band two patch columns wide (512 pixels in x) at each boundary. The final partition contained 1,092 training patches, 266 validation patches, 258 held-out test patches, and 107 excluded guard patches, which sum to 1,723. The split was spatial and not random.

Algorithm 1 Stage 2 controlled pipeline

Require: Retained patches $\mathcal{P}$ with bands B, G, R, NIR , indices NDVI, GNDVI, and mask m ; clustering seeds $S_c = \{42, 52, 62, 72, 82\}$; network seeds $S_n = \{42, 52, 62\}$

Ensure: Held-out metrics and paired bootstrap intervals

- 1: Assign every patch to training ($\mathcal{T}$), validation ($\mathcal{V}$), test ($\mathcal{E}$), or guard ($\mathcal{G}$) by image column; discard $\mathcal{G}$
- 2: Draw 500,000 vegetation pixels from $\mathcal{T}$; fit a standardizer on $Z_2 = [NDVI, GNDVI]$ using this sample only
- 3: **for** $k = 2, \dots, 6$ **do**
- 4: **for** each seed $s \in S_c$ **do**
- 5: Fit MiniBatch K-Means with k clusters; record the silhouette score
- 6: **end for**
- 7: Compute the pairwise ARI over all seed pairs
- 8: **end for**
- 9: Select k^* supported by both mean silhouette and mean ARI ($k^* = 2$)
- 10: Fit the k^* model with `random_state=42`; order clusters by mean NDVI as C1 and C2; freeze the standardizer and model
- 11: **for** each patch in $\mathcal{T} \cup \mathcal{V} \cup \mathcal{E}$ **do**
- 12: Set the reference label to background where $m = 0$; otherwise assign C1 or C2 with the frozen model
- 13: **end for**
- 14: Compute per-channel means and SDs from $\mathcal{T}$ only
- 15: Fit Logistic Regression on bands of $\mathcal{T}$ pixels using the reference labels
- 16: **for** each input configuration and each seed $s \in S_n$ **do**
- 17: Train U-Net on $\mathcal{T}$; keep the checkpoint with minimum loss on $\mathcal{V}$
- 18: **end for**
- 19: Predict $\mathcal{E}$ once for every method; aggregate confusion matrices
- 20: **for** $b = 1, \dots, 5000$ **do**
- 21: Draw $|\mathcal{E}| = 258$ test patches with replacement and use the same sampled patches for all methods; aggregate the confusion matrices per method
- 22: Compute A, macro-F1, and mIoU; store paired differences between methods
- 23: **end for**
- 24: Report the 2.5th and 97.5th percentiles of each difference

In **Figure 3**, each cell is one 256×256 -pixel patch, axes are PAN image coordinates in pixels and not map coordinates, and dashed lines mark the partition boundaries. The valid footprint is oblique, so the training region occupies the left and central part of the image, whereas the validation and test regions cover the right-hand part. Differences in class prevalence and in landscape among partitions are therefore confounded with geographic position. The validation region was used only for checkpoint selection. The held-out test region was not used for fitting the scaler or the clustering model, for input normalization, for checkpoint selection, or for choosing architectures or input configurations, and it entered only the final evaluation.

3.6.3 Cluster-number selection and clustering reference

Stage 2 clustering used a sample of 500,000 training-region vegetation pixels described by $Z_2 = [NDVI, GNDVI]$, standardized according to **Equation 15**,

$$z = \frac{x - \mu_{\mathcal{T}}}{\sigma_{\mathcal{T}}}, \quad (15)$$

where $\mu_{\mathcal{T}}$ and $\sigma_{\mathcal{T}}$ are the feature mean and standard deviation (SD) of the training sample. MiniBatch K-Means used `batch_size=4096` and `n_init=10`. For cluster-number selection, solutions with $k = 2, \dots, 6$ were fitted with five seeds (42, 52, 62, 72, and 82). Cluster separation was summarized by the mean and SD of the silhouette score [26] defined in **Equation 16**,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (16)$$

where $a(i)$ is the mean distance from sample i to the other members of its cluster and $b(i)$ is the smallest mean distance from sample i to the members of another cluster. Seed stability was summarized by the mean, SD, and range of the adjusted Rand index (ARI) [27] over the ten seed pairs. The value of k with the highest mean silhouette and the highest mean ARI was adopted as primary. Because a single value of k satisfied both criteria (Section 4.2), no tie-breaking rule was required. The value $k = 3$ appears in this analysis only as the Stage 1 setting, and no $k = 3$ model results are reported.

The clustering reference was then built by fitting the $k = 2$ model (`random_state=42`) to the same standardized sample. The fitted standardizer and clustering model were frozen and applied unchanged to the validation and test regions, and

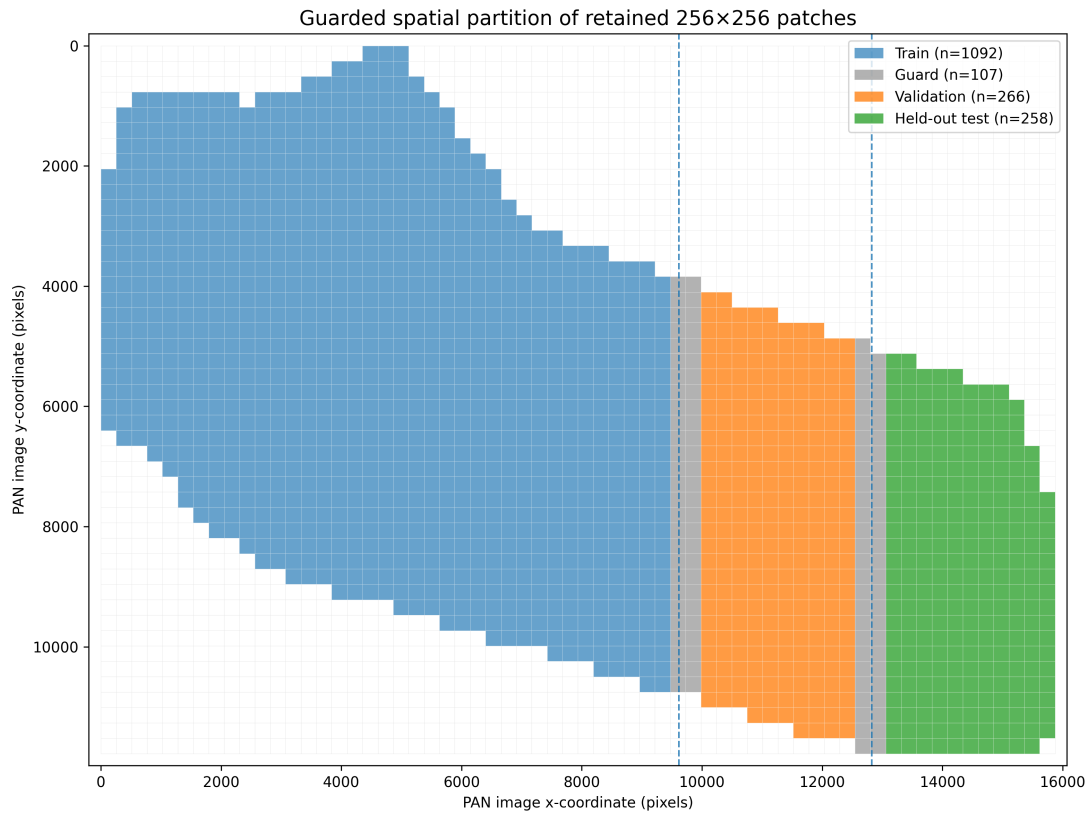


Figure 3: Guarded spatial partition of the retained patches

neither region contributed to fitting. Clusters were ordered by mean NDVI and labeled C1 (higher-NDVI stratum) and C2 (lower-NDVI stratum), and pixels outside the vegetation mask were labeled background. This frozen map is the reference for all Stage 2 comparisons.

3.6.4 Model inputs and training-only normalization

Three input configurations were defined: bands only (B, G, R, NIR), indices only (NDVI, GNDVI), and bands plus indices (B, G, R, NIR, NDVI, GNDVI). Because NDVI and GNDVI define the reference, the bands-only configuration is the leakage-safe test of whether the strata can be recovered without direct access to the clustering features. The other two configurations quantify the effect of target-feature overlap. All network inputs were normalized with per-channel means and SDs computed from the training patches only, as reported in **Table 6**, and validation and test data did not contribute to these statistics.

Table 6: Training-only normalization statistics

Statistic	B	G	R	NIR	NDVI	GNDVI
Mean	34.6769	47.9848	33.5449	195.1808	0.6578	0.5505
SD	8.8355	14.4004	13.3453	85.8382	0.1647	0.1610

3.6.5 Compared methods

The following methods were compared against the frozen reference.

- (A) *Direct frozen reference.* This is the pseudo-label definition itself, and its agreement with the reference is 1 by construction. It is not a competing predictive method.
- (B) *Bands-only Logistic Regression.* A non-spatial, pixel-wise classifier fitted to training-region pixels and the corresponding reference labels, with inputs *B, G, R, NIR* standardized by training statistics. The model used an L2

penalty with $C = 1.0$, the lbfgs solver, balanced class weights, a fitted intercept, $\text{max_iter}=500$, $\text{tol}=10^{-4}$, and $\text{random_state}=42$.

- (C) *U-Net variants*. The Stage 1 architecture family, a U-Net with an EfficientNet-B4 encoder without pretrained weights and three output classes, trained separately on each input configuration with seeds 42, 52, and 62.
- (D) 3×3 *majority smoothing*. Post-processing of the frozen reference, used only as a spatial-regularization diagnostic and not as an independent predictive model.

The U-Net variants shared the guarded split, training-only normalization, seeds, and augmentation. Training used the cross-entropy of **Equation 14** with $w_c = 1$ for all three classes, AdamW with learning rate 3×10^{-4} and weight decay 1×10^{-4} , a cosine schedule ($T_{\text{max}} = 30$), batch size 8, and at most 30 epochs. Training patches were augmented with random horizontal flips, vertical flips, and 90° rotations, each applied with probability 0.5. Only the checkpoint with minimum validation loss was retained, and the test region was used after checkpoint selection.

3.6.6 Evaluation metrics and spatial-block bootstrap

All Stage 2 methods were evaluated once on the held-out test region. Let n_{ij} be the number of test pixels with reference class i and predicted class j , with row sums $n_{i\cdot} = \sum_j n_{ij}$ and column sums $n_{\cdot j} = \sum_i n_{ij}$. Agreement, defined in **Equation 17**, and the class-wise scores, defined in **Equation 18**, were

$$A = \frac{1}{N} \sum_{p=1}^N \mathbf{1}[\hat{y}_p = y_p^{\text{ref}}], \quad (17)$$

$$F_{1,c} = \frac{2n_{cc}}{n_{c\cdot} + n_{\cdot c}}, \quad \text{IoU}_c = \frac{n_{cc}}{n_{c\cdot} + n_{\cdot c} - n_{cc}}, \quad (18)$$

$$\text{macro-F1} = \frac{1}{C} \sum_{c=1}^C F_{1,c}, \quad \text{mIoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c, \quad (19)$$

where y_p^{ref} is the reference label of pixel p , $\hat{y}_p$ the prediction, N the number of evaluated pixels, and $C = 3$ the number of classes (C1, C2, and background). All metrics were computed from confusion matrices aggregated over the evaluated region. Agreement is dominated by the prevalent C1 stratum, whereas macro-F1 and mIoU weight the three classes equally. Network results are reported as mean $\pm$ SD across the three seeds.

To respect spatial dependence among neighboring pixels [12], uncertainty was estimated with a block bootstrap in the spirit of [28], in which spatial blocks, not pixels, were the resampling units. Each of the 258 test patches was one block. In each replicate b , 258 blocks were drawn with replacement, the confusion matrices were aggregated, and the metrics of **Equations 17 to 19** were computed. For networks, each metric was computed separately per seed from the aggregated sampled-block matrix and then averaged across seeds. For a comparison of methods X and Y, the same sampled blocks were used for both, and the difference of **Equation 20** was computed for each of 5,000 replicates,

$$\Delta^{(b)} = m_X(\mathcal{B}_b) - m_Y(\mathcal{B}_b), \quad (20)$$

where m is a metric and $\mathcal{B}_b$ the sampled blocks. The 2.5th and 97.5th percentiles of $\Delta^{(b)}$ define the 95% interval, and a difference was regarded as statistically clear when this interval excluded zero. Because the analysis uses one scene, one date, and one spatial partition, these intervals quantify sampling uncertainty within the observed scene under an assumed block-independence model. The intervals do not establish practical importance or generalization to other farms, dates, sensors, or landscapes. Patch-wise averages of macro-F1 and mIoU were not used because classes absent from individual patches distort the averages.

3.7 Software, computational environment, and reproducibility

The GPU model and the full software versions of the original Stage 1 training were not recorded and cannot be reconstructed, so Stage 1 parameters are reported only where known. Stage 2 and the supplementary analyses ran in the documented environment of **Table 7**, which also collects the data dimensions, partition counts, and training settings needed to repeat the Stage 2 experiment. These versions are not necessarily those of the Stage 1 run.

Table 7: Stage 2 configuration and reproducibility summary

Item	Setting
Rasters	PAN 16026×11833 pixels; MSS 4008×2959 pixels; bilinear resampling of MSS to the PAN grid
Patch grid	256×256 pixels, stride 256; patches with less than 10% non-background pixels discarded; 1,723 retained
Partition	1,092 training, 266 validation, 258 held-out test, and 107 guard patches, split along the image column axis
Clustering	500,000 training-region vegetation pixels; NDVI and GNDVI standardized with training statistics; MiniBatch K-Means with <code>batch_size=4096</code> and <code>n_init=10</code> ; $k = 2, \dots, 6$ with seeds 42, 52, 62, 72, and 82; reference model with <code>random_state=42</code>
Selection of k	Highest mean silhouette and highest mean pairwise ARI; $k = 2$ satisfied both, so no tie-breaking was needed
Logistic Regression	L2 penalty, $C = 1.0$, <code>lbfgs</code> , balanced class weights, <code>max_iter=500</code> , <code>tol=10⁻⁴</code> , <code>random_state=42</code>
Networks	U-Net with EfficientNet-B4 encoder without pretrained weights; seeds 42, 52, and 62; AdamW, learning rate 3×10^{-4} , weight decay 1×10^{-4} ; cosine schedule ($T_{\max} = 30$); batch size 8; at most 30 epochs; flips and 90° rotations with probability 0.5; checkpoint with minimum validation loss
Bootstrap	258 test patches as blocks; 5,000 paired replicates; percentile interval (2.5th and 97.5th)
Software	Python 3.12.12, NumPy 2.0.2, Rasterio 1.5.0, scikit-learn 1.6.1, PyTorch 2.10.0+cu128, <code>segmentation-models-pytorch</code> 0.5.0, Albumentations 2.0.8, SciPy 1.16.3, scikit-image 0.25.2, CUDA 12.8
Hardware	Tesla T4 GPU

4 RESULTS

4.1 Stage 1 exploratory proof of concept

Over the 22,609,920 pixels of the 345 validation patches that had also been used for checkpoint selection, the saved Stage 1 U-Net reproduced its four-class $k = 3$ pseudo-labels with agreement 0.9806, macro-F1 0.9830, mIoU 0.9667, and balanced accuracy 0.9843. These historical exploratory results are not used as evidence of independent generalization, of the need for a deep network, or of any biological property, for the reasons given in Section 3.5. The controlled results below supersede them.

A supplementary audit of the Stage 1 configuration removed the duplicated GNC feature and re-clustered with [NDVI, GNDVI, RVI] at $k = 3$. The result had an ARI of only 0.4429 against the original four-feature clustering, with 76.9884% ordered-stratum agreement. The Stage 1 strata were therefore materially sensitive to feature duplication, which motivated the redundancy audit and the cluster-number analysis of Stage 2. This audit was diagnostic and historical, and it did not use the final Stage 2 feature set.

4.2 Feature redundancy and cluster-number selection

Figure 4a shows the training-region feature audit. The Spearman correlation between NDVI and RVI was 0.9999999973, the maximum absolute deviation between NDVI and $f(RVI)$ was 1.7881393×10^{-7} , and the mean deviation was 2.907607×10^{-8} . RVI was an algebraic re-expression of NDVI and was removed, as was GNC. The three visible bands were highly correlated with one another (0.90 to 0.96, including 0.96097 for blue and red), but the bands are not algebraically related, and because correlation alone was not a removal criterion, all four bands remained inputs to the bands-only models. NDVI and GNDVI were correlated at 0.95 without being algebraically dependent. NDVI correlated strongly with the NIR band (0.80) and hardly at all with the red band (0.03), which indicates that NIR variation dominates the NDVI ranking in this scene. The clustering representation was therefore NDVI and GNDVI.

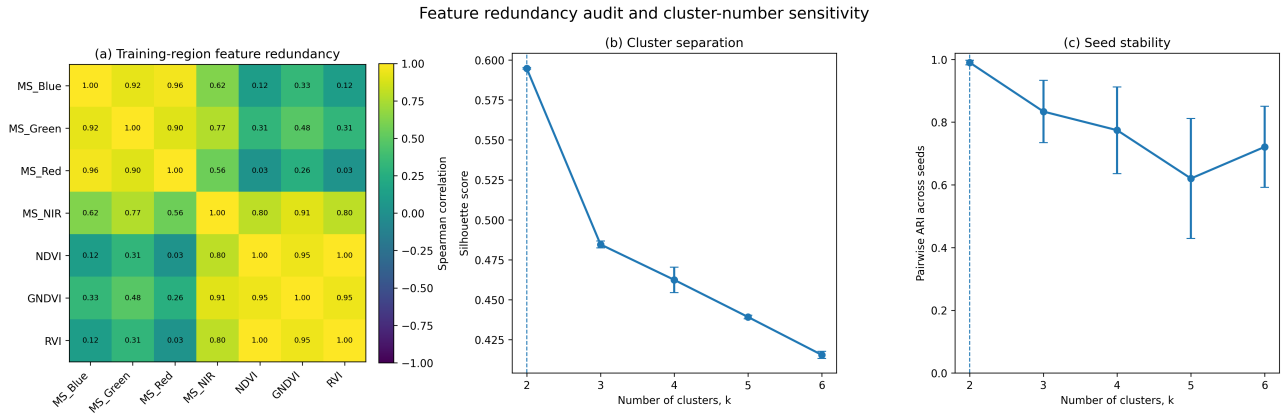


Figure 4: Feature redundancy audit and cluster-number sensitivity in the training region

The five-seed analysis favored $k = 2$ on both criteria, as shown in **Table 8** and **Figures 4b and 4c**. Panels (b) and (c) show means as points and SDs across seeds as error bars, with a dashed line at $k = 2$. The two-cluster solution had the highest mean silhouette (0.59472 ± 0.00048) and the highest pairwise ARI (0.99034 ± 0.00723 , range 0.97837 to 0.99818). For $k = 3$, the silhouette fell to 0.48460 ± 0.00223 and the ARI to 0.83385 ± 0.09944 (range 0.68544 to 0.96170), which indicates seed-dependent partitions. The silhouette continued to decline for $k = 4$ to 6. The ARI was lowest at $k = 5$ (0.62038) and rose slightly at $k = 6$ (0.72116), so stability was not monotonic in k . This is expected because ARI measures partition agreement across seeds and not cluster quality, and no larger k approached the stability of $k = 2$. The two-stratum configuration was adopted. Both criteria tend to favor few clusters when a feature distribution is continuous, so the result shows that a two-way split is reproducible and not that the scene contains two natural classes.

Table 8: Cluster-number sensitivity in the training region

k	Silhouette	Pairwise ARI
2	0.59472 ± 0.00048	0.99034 ± 0.00723
3	0.48460 ± 0.00223	0.83385 ± 0.09944
4	0.46241 ± 0.00795	0.77433 ± 0.13833
5	0.43917 ± 0.00112	0.62038 ± 0.19154
6	0.41544 ± 0.00222	0.72116 ± 0.12946

4.3 Spatially controlled clustering reference

The guarded partition of **Figure 3** yielded 1,092 training, 266 validation, and 258 held-out test patches, with 107 guard patches excluded. In the 500,000-pixel training sample, the higher-NDVI cluster (C1) contained 339,082 pixels with mean NDVI 0.74041 and mean GNDVI 0.64074, and the lower-NDVI cluster (C2) contained 160,918 pixels with mean NDVI 0.56854 and mean GNDVI 0.43135.

Table 9 gives the reference label distribution by partition. C1 was the majority stratum in every partition (65.126% to 73.685%), followed by C2 (23.438% to 30.808%) and background (2.876% to 4.065%). Prevalence differs among partitions because the partitions are geographically separated, which is why class-balanced metrics are reported alongside agreement. The proportions describe retained patches and not physical scene composition.

Table 9: Distribution of the frozen clustering reference by partition

Partition	Patches	C1 (higher NDVI)	C2 (lower NDVI)	Background
Training	1,092	46,607,839 (65.126%)	22,048,083 (30.808%)	2,909,390 (4.065%)
Validation	266	12,845,208 (73.685%)	4,085,927 (23.438%)	501,441 (2.876%)
Held-out test	258	11,945,088 (70.646%)	4,299,165 (25.426%)	664,035 (3.927%)

4.4 Controlled model comparison

Every controlled model reproduced the reference in the held-out test region with agreement above 0.99, as reported in **Table 10** and illustrated in **Figure 5**. The bands-only Logistic Regression has no access to NDVI or GNDVI and reached agreement 0.994007, macro-F1 0.989465, and mIoU 0.979209. The bands-only U-Net reached agreement 0.991922 ± 0.001526 , numerically below Logistic Regression, with higher macro-F1 (0.991909 ± 0.001372) and mIoU (0.984010 ± 0.002674); this pair was not part of the paired bootstrap comparisons.

Table 10: Held-out test reproduction of the frozen clustering reference

Method	Inputs	Agreement	Macro-F1	mIoU
Logistic Regression	Bands	0.994007	0.989465	0.979209
U-Net	Bands	0.991922 ± 0.001526	0.991909 ± 0.001372	0.984010 ± 0.002674
U-Net	Bands + indices	0.993406 ± 0.000693	0.993202 ± 0.000905	0.986534 ± 0.001782
U-Net	Indices	0.994176 ± 0.000403	0.994146 ± 0.000326	0.988391 ± 0.000639

Table 10 presents networks as mean $\pm$ SD over seeds 42, 52, and 62, and Logistic Regression as a single value. Bands are $[B, G, R, NIR]$ and indices are $[NDVI, GNDVI]$. Agreement is the overall pixel-wise accuracy, and macro-F1 and mIoU are averaged over the classes. The rows below the rule include the clustering features as inputs and are target-feature overlap ablations. Among U-Net variants, adding the clustering indices raised every metric. The indices-only values (agreement 0.994176 ± 0.000403 , macro-F1 0.994146 ± 0.000326 , mIoU 0.988391 ± 0.000639) were numerically above those of the bands-only U-Net. This is not a clean comparison, because NDVI and GNDVI define the reference, and the two index-based rows are interpreted in Section 4.5 as evidence of target-feature overlap, not as the best leakage-safe performance.

Figure 5 shows agreement, macro-F1, and mIoU for the four configurations, with network bars as means across three seeds, error bars of ± 1 SD, and Logistic Regression as a single value. The y-axis is truncated at 0.975, which visually magnifies differences that are small in absolute terms. All bars exceed 0.979. The gap between agreement and mIoU is largest for Logistic Regression (about 0.015) and smaller for the bands-only U-Net (about 0.008), which is consistent with the U-Net improving mainly the class-balanced and overlap-based scores in point estimate. The bands-only U-Net shows the widest error bars. The smoothing diagnostic is not plotted because it modifies the reference map directly.

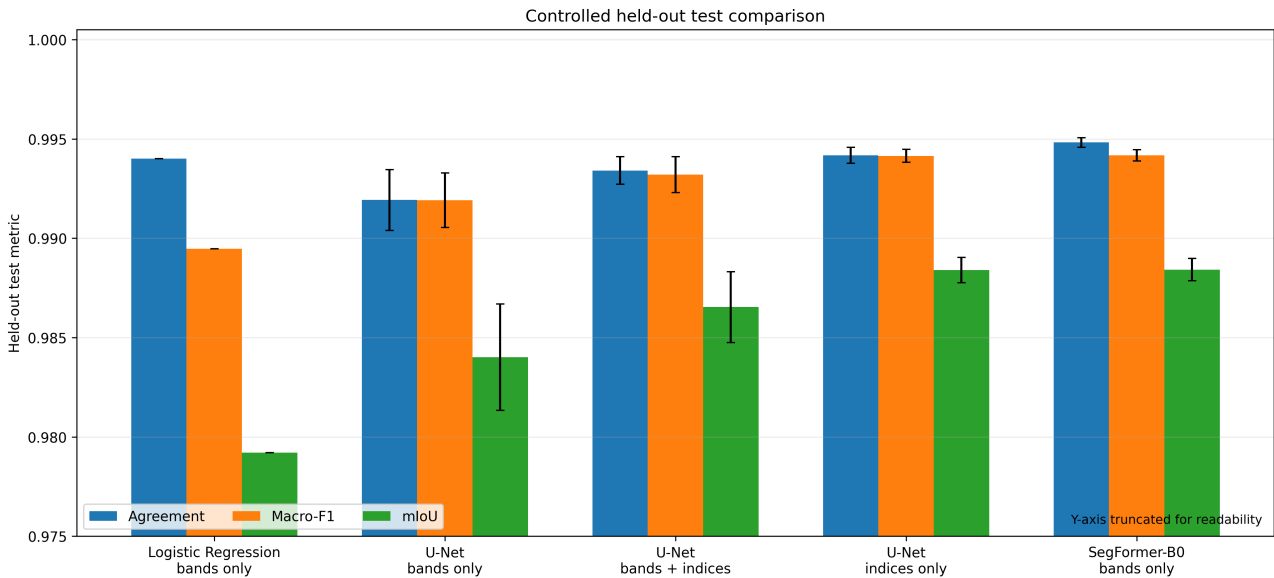


Figure 5: Held-out test comparison of models on the frozen clustering reference

As a separate diagnostic, 3×3 majority smoothing of the frozen reference agreed with the unsmoothed reference at 0.999553 (macro-F1 0.999601, mIoU 0.999203), changing about 0.045% of held-out test labels. Simple local post-processing therefore removes small isolated regions while altering very few labels.

4.5 Spatial-block uncertainty and paired comparisons

Table 11 reports paired differences and 95% intervals from the spatial-block bootstrap.

Table 11: Paired spatial-block bootstrap differences on the held-out test region

Comparison	Metric	Difference	95% CI	CI excludes 0
U-Net indices only vs. U-Net bands only	Agreement	+0.002253	[0.001690, 0.002839]	Yes
	Macro-F1	+0.002237	[0.001564, 0.003077]	Yes
	mIoU	+0.004381	[0.003060, 0.006123]	Yes
U-Net bands + indices vs. U-Net bands only	Agreement	+0.001483	[0.001144, 0.001832]	Yes
	Macro-F1	+0.001293	[0.000971, 0.001630]	Yes
	mIoU	+0.002524	[0.001873, 0.003177]	Yes

The table is based on 258 spatial blocks and 5,000 paired replicates with identical sampled blocks for both methods, and differences are the first method minus the second. Intervals are percentile intervals without bias correction.

Supplying the clustering indices to the U-Net produced small but consistent gains over the bands-only U-Net. For indices only, the differences were +0.002253 in agreement (0.001690 to 0.002839), +0.002237 in macro-F1 (0.001564 to 0.003077), and +0.004381 in mIoU (0.003060 to 0.006123). For bands plus indices, the differences were +0.001483 (0.001144 to 0.001832), +0.001293 (0.000971 to 0.001630), and +0.002524 (0.001873 to 0.003177). Direct access to the features that define the reference measurably increased reproduction, which confirms that target-feature overlap contributed to the apparent performance of the original eight-channel design.

4.6 Qualitative held-out examples

Figure 6 shows two held-out test blocks selected algorithmically. Using block-level pixel agreement of the seed-42 network predictions, the first example (patch 1791) is the block closest to the median and the second (patch 2593) the block closest to the 10th percentile. Columns show an RGB composite, the frozen reference, and the bands-only Logistic Regression and U-Net predictions, with the seed-42 runs for networks. Green is C1, pale yellow is C2, and dark would denote background, which is absent from both blocks.

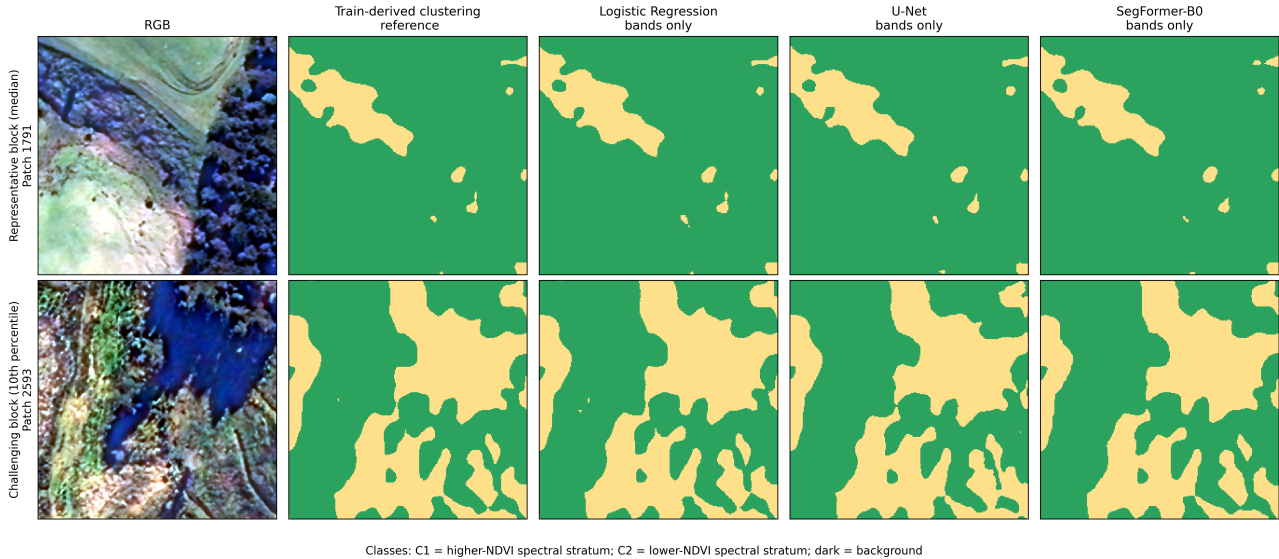


Figure 6: Qualitative held-out test examples

In the median block, C1 dominates, and the C2 regions form a diagonal band at upper left plus a few small fragments. Both bands-only models reproduce these structures, and differences are limited to outlines and a few small fragments. In the 10th-percentile block, the reference is more fragmented and C2 covers a large area at upper right and lower center. In the RGB composite this area broadly coincides with a large dark blue region, which suggests that low NDVI there is linked to shaded or dark surfaces, although the figure cannot confirm the surface type. The models differ mainly along

stratum boundaries. The reference and the Logistic Regression maps retain a few isolated specks that the U-Net map largely removes. These examples illustrate the quantitative results and do not replace them.

4.7 Reduced-resolution assessment of Brovey preprocessing

Across 6,835,681 valid evaluation pixels, the bilinear no-fusion baseline and Brovey fusion both produced a SAM of 1.1769° , whereas ERGAS was 1.4333 for the baseline and 21.3025 for Brovey, as reported in **Table 12**. The identical SAM follows from the common multiplicative factor of the idealized Brovey transform in **Equation 4**, since SAM depends on band ratios and not on absolute magnitude, and invariance to a common multiplicative scaling is a known property of SAM [21]. The much larger ERGAS for Brovey reflects the rescaling of band values toward PAN intensity in **Equation 2**, which changes the absolute radiometry that ERGAS penalizes. Brovey therefore did not improve spectral or radiometric fidelity in this experiment.

Table 12: Reduced-resolution fusion quality

Method	SAM (degrees)	ERGAS
Bilinear no-fusion baseline	1.1769	1.4333
Brovey fusion	1.1769	21.3025

5 DISCUSSION

5.1 Strong and reproducible scene-relative spectral structure

Three results jointly answer RQ1 to RQ3. After removal of the exact (GNC) and algebraic (RVI) duplicates, the NDVI and GNDVI feature space supported a two-stratum partition that was clearly separated (silhouette 0.5947) and nearly invariant to the seed (pairwise ARI 0.9903). The partition fitted only in the training region and then frozen remained well populated in a guarded, geographically separated test region, and every bands-only model reproduced it there with agreement above 0.99. The scene therefore contains strong, structured, and learnable scene-relative spectral information, and raw bands suffice to reproduce most of the clustering structure outside the region where it was defined.

For RQ2, the supported configuration is $k = 2$, not the $k = 3$ used in Stage 1. The three-level stratification in Stage 1 was an operational choice and not a stable property of the scene, consistent with its sensitivity to feature duplication in Section 4.1. As noted in Section 4.2, silhouette and ARI favor few clusters on continuous data, so the partition should be regarded as a reproducible split of a spectral continuum. Reproducibility is used here in a restricted sense, namely across clustering seeds and across space within one scene and one date.

5.2 High pseudo-label reproduction is not unique to deep learning

The bands-only Logistic Regression reached 0.9940 agreement, numerically above the bands-only U-Net mean (0.9919 ± 0.0015) and above the bands-only U-Net in agreement. This outcome follows from how the reference is built. A level set $\text{NDVI} = t$ satisfies $(1 - t) \text{NIR} = (1 + t) R$, a plane through the origin in band space, and the same holds for GNDVI. The background rule of **Equation 13** combines two conditions that are also linear in the bands. The remaining boundary between C1 and C2 is a line in standardized NDVI and GNDVI space and hence a smooth surface in band space, which a linear classifier can approximate over the limited range of values present in the scene. A pixel-wise classifier without spatial context can thus recover the reference closely. The high agreement of the Stage 1 U-Net (0.9806) accordingly does not show that a deep network was necessary. It shows that the pipeline runs end to end and that most of the pseudo-label structure is available from per-pixel spectra.

5.3 Value of spatial modeling and of the architecture

The bands-only U-Net did not exceed Logistic Regression in overall agreement, whereas its macro-F1 and mIoU were higher in point estimate. This pair was not part of the paired bootstrap, so the macro-F1 and mIoU differences cannot be called statistically clear. Because agreement is dominated by the prevalent C1 stratum while macro-F1 and mIoU weight the classes equally, the pattern is consistent with gains concentrated in less prevalent classes, although per-class scores were not tabulated and this cannot be confirmed.

The spatial benefit of the U-Net was modest. The bands-only U-Net had the largest seed-to-seed SD of the network configurations. All point-estimate differences among bands-only models were below 0.01. The reference is spectrally simple and contains few isolated pixels, since 3×3 majority smoothing altered only about 0.045% of test labels, so the scope for spatial regularization to change overall agreement is small. Class-balanced and overlap-based metrics are the more informative indicators. The answer to RQ4 is that the U-Net showed no benefit in overall agreement and a small point-estimate benefit in those metrics that was not tested statistically. This result holds for the fixed training budget used here, without hyperparameter tuning for any model.

5.4 Target-feature overlap

Adding NDVI and GNDVI to the U-Net inputs produced small but clear increases in all three metrics relative to bands-only input, which confirms that target-feature overlap contributed to the apparent performance of the original eight-channel U-Net. Because Brovey fusion degraded reduced-resolution radiometric fidelity, these classification results are conditional on that preprocessing choice, and no classification run without fusion was performed. The effect is small in absolute terms because the bands already contain the information from which the indices are computed. It nevertheless shows that agreement can be raised by supplying the variables that define the targets, which carries no biological meaning. The leakage-safe conclusions therefore rest on the bands-only comparisons.

5.5 Implications of Brovey preprocessing

The cancellation in **Equation 4** and the identical SAM values show that Brovey fusion enlarged the sampling grid without creating independent sub-meter multispectral measurements. The much larger ERGAS (21.3025 versus 1.4333) shows that Brovey degraded radiometric fidelity in the reduced-resolution assessment. Brovey was retained to keep Stage 2 comparable with Stage 1, and it should not be regarded as a preferred preprocessing strategy for future palm-monitoring applications. A no-fusion baseline was evaluated for fusion fidelity only, and repeating the main classification experiment on the resampled MSS bands without fusion is a priority for future work. Methods that balance spectral and spatial fidelity [14, 15, 16] should also be evaluated, particularly if per-palm spectral attributes are to be calibrated against field measurements.

5.6 Implications for a future individual-palm monitoring campaign

Eventual individual-palm delineation, field-calibrated assessment of plant water status or stress, and transfer across farms require two prerequisites. The first is technical, namely that the imagery must contain strong, structured, and learnable spectral information, and the present results establish this for one scene and date. The second is biological, namely that the image-derived strata, or quantities derived from the imagery, must correspond to specific physiological states. Nothing in the experiment shows that C1 and C2 correspond to differences in moisture, water status, stress, disease, or palm condition.

The answer to RQ5 is therefore that the evidence suffices to motivate, but not to replace, a subsequent study. That study should combine imagery from several farms and dates, individual-palm boundaries annotated for delineation, field measurements of moisture, water status, or stress for individual palms, and external testing on farms withheld from training with spatially or farm-blocked validation [12, 13]. Comparison across dates and farms will also require radiometric calibration, atmospheric correction, and RPC-based orthorectification, none of which was applied here. Whether the 0.5 m PAN sampling and approximately 2 m multispectral sampling suffice for individual-palm delineation and per-palm spectral attribution was not tested and must be established with annotated crowns. The C1 and C2 strata could stratify field sampling so that measurements span the observed spectral range, and whether the strata align with measured plant water status then becomes an empirical question for that study.

6 LIMITATIONS AND FUTURE WORK

Several limitations bound the conclusions, and the most important concern the evaluation design. Only one guarded left-to-right partition was evaluated, so partition is confounded with position in the scene, and the bootstrap intervals describe sampling variability within that single partition. Per-class results and an alternative partition are the first additions required. The remaining limitations are grouped below. The first group concerns the data. Only one Jilin-1 scene from one acquisition date was analyzed, which prevents any assessment of temporal persistence, and no external farm or scene was available for transfer testing. No independent biological or physiological ground truth and no individual-palm annotations existed, and the scene lies in New Zealand, where palm occurrence was not established. The strata therefore cannot be read as moisture, stress, disease, or health classes.

The second group concerns the reference. The strata are outputs of a clustering model fitted to two correlated indices. The choice of $k = 2$ rests on silhouette and ARI, which favor small cluster counts for continuous distributions, so the partition may subdivide a continuum. The reference also depends on the vegetation-mask thresholds and on the feature set, as the Stage 1 sensitivity to duplicated features showed.

The third group concerns the evaluation design. Only one guarded partition and one patch size (256×256 pixels) were evaluated. The guard reduces but cannot eliminate spatial dependence within a single scene, and the left-to-right arrangement confounds partition with position in the scene. The block bootstrap treats test patches as independent units although neighboring patches may be correlated, so the intervals may be optimistic, because 258 adjacent patches are unlikely to be fully independent, and a sensitivity analysis with larger blocks (for example, 2×2 or 3×3 patch groups) was not performed. Per-class precision, recall, F1, and IoU were not tabulated, so the class responsible for the macro-F1 and mIoU difference between the U-Net and Logistic Regression is not identified. Network results rest on three seeds, and all models were trained with a fixed budget and no hyperparameter search, so the comparison is conditional on that budget. The U-Net was trained from scratch, and pretrained weights were not examined.

The fourth group concerns preprocessing. Brovey fusion did not improve reduced-resolution fidelity and had a much larger ERGAS than the no-fusion baseline, and the main classification experiment was not repeated without fusion. PAN and MSS registration error was not quantified, and neither radiometric calibration, atmospheric correction, nor orthorectification was applied, so the analysis stayed in image coordinates with uncalibrated DN values.

The final group concerns reproducibility. The Stage 1 training software versions are not fully recoverable, although the Stage 2 environment is documented in **Table 7**. Redistribution of the commercial imagery may be restricted by licensing, and the analysis code is not publicly released at present. Future work should address these points through multi-date, multi-farm imagery with field-validated labels and palm annotations, calibrated radiometry, repeated spatial partitions with several patch sizes and block sizes, a no-fusion classification baseline, and per-class reporting.

7 CONCLUSION

This feasibility study asked whether a single Jilin-1 scene contains enough reproducible spectral-spatial structure to justify a multi-farm, field-validated individual-palm monitoring campaign. The Stage 1 proof of concept produced high same-scene agreement (0.9806), but feature redundancy, whole-scene clustering, target-feature overlap, and validation reuse prevented it from answering that question. The controlled Stage 2 experiment addressed these dependencies.

The experiment shows that the scene contains strong, reproducible, scene-relative spectral structure that raw multispectral bands recover in a geographically held-out region. After removal of GNC and RVI as duplicates, a two-stratum NDVI and GNDVI partition was well separated and nearly seed-invariant (silhouette 0.5947, ARI 0.9903). Strata defined only in the training region were reproduced in a guarded held-out test region. High pseudo-label reproduction was not unique to the U-Net, because a bands-only Logistic Regression reached comparable overall agreement (0.9940). The bands-only U-Net did not exceed Logistic Regression in overall agreement. Supplying the clustering indices to the U-Net produced small but consistent gains, which confirms that target-feature overlap contributed to the original apparent performance. Because Brovey fusion degraded reduced-resolution radiometric fidelity, these results are conditional on that preprocessing choice, and a classification run without fusion remains to be performed.

These results do not constitute validated detection of plant stress, moisture deficit, disease, or palm condition, and palm presence in the analyzed scene was not verified. The findings provide technical support for the next research phase, which requires multi-farm and multi-date imagery, individual-palm annotation, field measurements of moisture and stress, independent biological validation, and cross-farm transfer testing.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Not Applicable.

DATA AVAILABILITY

This study used commercially supplied Jilin-1 imagery. The dataset is not publicly available because of applicable commercial and licensing restrictions but may be obtained from the corresponding author upon reasonable request, subject to the terms of the relevant data-use agreement.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the use of OpenAI's ChatGPT for assistance with manuscript refinement.

FUNDING

No Funding.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

REFERENCES

- [1] R. Al-Ruzouq, M. B. A. Gibril, A. Shanableh, J. Bolcek, F. Lamghari, N. A. Hammour, A. El-Keblawy, and R. Jena, "Spectral-spatial transformer-based semantic segmentation for large-scale mapping of individual date palm trees using very high-resolution satellite data," *Ecological Indicators*, vol. 163, p. 112110, 2024.
- [2] M. B. A. Gibril, H. Z. M. Shafri, A. Shanableh, R. Al-Ruzouq, S. J. bin Hashim, A. Wayayok, and M. S. Sachit, "Large-scale assessment of date palm plantations based on uav remote sensing and multiscale vision transformer," *Remote Sensing Applications: Society and Environment*, vol. 34, p. 101195, 2024.
- [3] A. Hammadi, I. Mahi, H. Rhinane, M. Maanan, A. Zingraff-Hamed, and M. Maanan, "Deep learning for palm tree health assessment: UAV-based segmentation in the figuig region of morocco," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4/W17-2025, pp. 169–181, 2026.
- [4] S. Delalieux, T. Hardy, M. Ferry, S. Gomez, L. Kooistra, M. Culman, and L. Tits, "Red palm weevil detection in date palm using temporal uav imagery," *Remote Sensing*, vol. 15, no. 5, 2023.
- [5] J. Rouse, John W., R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the great plains with ERTS," in *Proceedings of the Third Earth Resources Technology Satellite-1 Symposium*, vol. 1, (Washington, DC), pp. 309–317, National Aeronautics and Space Administration, 1974. NASA Special Publication SP-351.
- [6] C. F. Jordan, "Derivation of leaf-area index from quality of light on the forest floor," *Ecology*, vol. 50, no. 4, pp. 663–666, 1969.
- [7] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzlyak, "Use of a green channel in remote sensing of global vegetation from eos-modis," *Remote Sensing of Environment*, vol. 58, no. 3, pp. 289–298, 1996.
- [8] S. K. McFeeters, "The use of the normalized difference water index (ndwi) in the delineation of open water features," *International journal of remote sensing*, vol. 17, no. 7, pp. 1425–1432, 1996.
- [9] X. Lu, Z. Jiang, and H. Zhang, "Weakly supervised remote sensing image semantic segmentation with pseudo-label noise suppression," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.
- [10] Z. Wang, H. Chang, L. Bai, L. Chen, and X. Bi, "A creative weak supervised semantic segmentation for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

- [11] X. Xue, H. Zhu, X. Li, J. Wang, L. Qu, and B. Hou, “Egpo: Enhanced guidance and pseudo-label optimization for semi-supervised semantic segmentation of remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.
- [12] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.
- [13] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, “blockcv: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models,” *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 225–232, 2019.
- [14] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, “A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 53–81, 2021.
- [15] M. Ciotola, G. Guarino, and G. Scarpa, “An unsupervised cnn-based pansharpening framework with spectral-spatial fidelity balance,” *Remote Sensing*, vol. 16, no. 16, 2024.
- [16] W. Kang, Y. Feng, Y. Ding, H. Xiang, X. Liu, and Y. Cai, “Hsstn: A hybrid spectral–structural transformer network for high-fidelity pansharpening,” *Remote Sensing*, vol. 17, no. 19, 2025.
- [17] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.
- [18] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105–6114, PMLR, 2019.
- [19] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [20] N. A. A. Braham, C. M. Albrecht, J. Mairal, J. Chanussot, Y. Wang, and X. X. Zhu, “Spectraearth: Training hyperspectral foundation models at scale,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 16780–16797, 2025.
- [21] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, “The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data,” *Remote Sensing of Environment*, vol. 44, no. 2, pp. 145–163, 1993. Airbone Imaging Spectrometry.
- [22] L. Wald, “Quality of high resolution synthesised images: Is there a simple criterion?,” in *Proceedings* (T. Ranchin and L. Wald, eds.), (Sophia Antipolis, France), pp. 99–103, SEE/URISCA, 2000.
- [23] D. Sculley, “Web-scale k-means clustering,” in *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, ACM, 2010.
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [25] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.
- [26] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
- [27] L. J. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, pp. 193–218, 1985.
- [28] H. R. Künsch, “The jackknife and the bootstrap for general stationary observations,” *The Annals of Statistics*, vol. 17, no. 3, pp. 1217–1241, 1989.