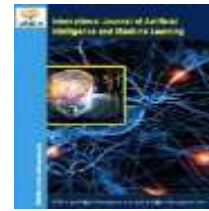




Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Interpretable Patch Screening and Guided Attention U-Net For Floating Marine Debris Detection

Swati B. Magare¹, Ratnadeep R. Deshmukh², Ranjana Gore³, Priyanka U. Randive⁴

^[1] Research Scholar, Dr. BAMU Aurangabad orcid : 0009-0003-5516-5830

^[2] Dr. Ratnadeep R. Deshmukh, Professor, Dr. BAMU, Aurangabad

^[3] Assistant Professor, MIT Art Design and Technology University Pune, Loni Kalbhori, Pune ORCID: 0000-0002-6934-274

^[4] Research fellow Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad

^[1] sbmagare56@gmail.com, ^[2] rrdeshmukh.csit@bamu.ac.in, ^[3] goreranjana123@gmail.com, ^[4] priyankarandive7@gmail.com

Abstract

Floating marine debris is spectrally ambiguous and represents only a small fraction of Sentinel-2 imagery. In this work, we present two complementary debris-aware implementations, which are evaluated with the official partitions of the Marine Debris Archive (MARIDA). First, we propose an interpretable patch-screening baseline that translates eleven-band patches to 54 spectral, index and texture descriptors and compares logistic regression, radial-basis support vector machines and Random Forests. On 359 held-out patches, the Random Forest achieves 0.774 accuracy, 0.585 precision, 0.683 recall, 0.630 F1, 0.821 ROC-AUC and 0.640 average precision. Second, the proposed approach introduces a 350,414-parameter Physics-Guided Spectral-Spatial Attention U-Net for sparse pixel segmentation. It augments the eleven spectral bands with differentiable NDVI, NDWI, Floating Debris Index and Plastic Index channels, learns patch-conditioned spectral gates, applies multiscale convolution and bottleneck self-attention, and is optimised with hard-negative binary cross-entropy, recall-oriented Tversky loss and an edge term. Only on validation data the threshold of 0.75, two-pixel component filter and morphological closing are selected. On the locked test set, the Proposed method achieves 0.983 ROC-AUC, 0.713 average precision, 0.201 precision, 0.530 recall, 0.291 F1 and 0.170 IoU.

Keywords: marine debris; Sentinel-2; MARIDA; AquaSpectra; physics-guided learning; spectral attention; semantic segmentation; reproducibility.

1. Introduction

Marine litter is a persistent environmental pressure with sources, transport pathways and accumulation areas across rivers, coasts and open waters. Conventional observations from ships, beaches and aircraft are valuable but expensive to repeat over large areas. Earth observation therefore provides an attractive complementary perspective, with systematic acquisitions that can assist retrospective analyses, event screening and prioritisation of field surveys. But the phrase 'satellite detection of plastic' can exaggerate what a multispectral sensor actually measures. At ten-to-twenty metre ground sampling, a Sentinel-2 pixel is generally a mixture of water, illumination effects, atmosphere, natural floating matter and possibly debris. The task that can be defended scientifically is thus to identify an anomalous spectral-textural signature associated with verified marine-debris events, not to infer material composition without corroboration [1], [2].

Sentinel-2 is particularly relevant because its MultiSpectral Instrument measures visible, red-edge, near-infrared and short-wave-infrared wavelengths at a useful revisit frequency. Various aspects of the water-object contrast are described by band combinations such as the Floating Debris Index, Normalised Difference Vegetation Index and Normalised Difference Water Index [2],[3]. Floating material can increase near-infrared reflectance relative to the surrounding water. Experiments conducted as part of the Controlled Plastic Litter Project showed that artificial floating targets can be detected under appropriate coverage and sea-state conditions, but also suggested how target fraction, geometry, atmosphere, waves and adjacency affect the measured signal [3],[4]. These results provide physical meaning to feature engineering, but also caution against universal thresholds. To move the field from isolated case studies to comparative machine learning the MARIDA benchmark has been proposed [1]. It contains eleven-band Sentinel-2 patches, sparse pixel annotations, and patch-level multi-label metadata for marine debris and confounding classes, including natural organic material, ships, waves, wakes, foam,

turbid water, shallow water, and clouds. Crucially, its published partitions contain all patches from a scene and date in one subset. This design reduces the spatial-temporal debris and allows a more realistic estimate of transfer to unseen scenes [1]. Any implementation that does not respect these partitions, that samples folders in OS order, or that maps all annotated pixels to the debris class is subverting the purpose of the benchmark. The reviewed implementation brought together several promising ideas at first – spectral indices, spectral statistics, texture features, a CNN branch, attention-like weighting and a web interface. But the branches did not make up a consistent train-to-inference system. The CNN was initialised randomly and never trained. Only 6 features were used for app inference while other components provided 118 features. The band positions of Sentinel-2 were incorrect for MARIDA. The binary labels were based on any non-zero value in the class-mask. Also, comparison plots used hard-coded external scores and the app looked for a ground-truth file next to a random temporary upload path. These are not cosmetic defects each of which can materially change reported performance or create an evaluation that cannot be reproduced.

Marine Debris is the positive class and all other MARIDA patch is the comparison class. The original official splits are kept: 694 training patches, 328 validation patches, and 359 test patches. The image representation consists of 44 per-band distribution descriptors, six summary statistics of three normalized-difference indices, and four GLCM texture measures. We compare three conventional learners in class-aware training. We choose a probability threshold on validation F1 and then fix it for the test set. The separation between fitting, selection and final evaluation is important because class imbalance makes the default 0.5 threshold arbitrary and accuracy alone is uninformative.

First, the study establishes an interpretable baseline, as a valid deep experiment needs trained weights, controlled augmentation, pixel level losses and an explicit selection protocol. Then it presents Proposed methodology as a stand-alone extension of trained segmentation. The ordering rules out the possibility that architectural novelty hides data debris or label errors: the compact Random Forest shows the corrected benchmark behaviour, and the physics-guided spectral-spatial network tests if learned multiscale context can localise sparse annotated pixels. Neither task is considered a substitute for the other, and published segmentation results [5]-[8] are only discussed when the task and protocol are compatible.

There are five types of contribution:

First, the implementation is fixed end to end, including band mapping, normalisation, label semantics, official splitting, model persistence and upload safety.

Second, a reproducible Random Forest provides an interpretable patch-level reference.

Third, the proposed methodology introduces a novel trained pixel-level architecture that combines differentiable physics indices, learned spectral gating, multiscale spatial encoding, and attention.

Fourth, all thresholds and settings for morphology are chosen on validation data, and frozen prior to locked-test evaluation.

Fifth, the Streamlit laboratory shows all processing steps and all reported tables and graphs are produced from saved experimental artefacts rather than manually entered values.

The rest of the paper reviews spectral and learning-based marine-debris detection, describes the proposed workflow, presents reproducible results, and discusses limitations and implications for deployment. In the manuscript, the word 'detection' refers to the screening of patch-level evidence. It doesn't mean a debris inventory confirmed by materials or pixel perfect. This careful scope allows a reasonable matching of the claimed result with the available annotations and allows responsible usage in upcoming multi-stage monitoring systems.

2. Literature Review

Early satellite studies have shown that dense floating aggregations perturb the otherwise dark near-infrared response of water. Biermann et al. applied a Floating Debris Index combined with spectral-shape classification and reported promising discrimination for selected coastal cases [2]. Later controlled-target studies related detectability with the fraction of plastic in a mixed Sentinel-2 pixel and showed that favourable geometry and environmental conditions are decisive [3],[4]. Themistocleous et al. also investigated a Plastic Index and reported the near-infrared contrast [9]. The studies together gave rise to spectral arithmetic, but did not resolve the problem of confounding by vegetation, foam, sediment or mixed natural debris.

The creation of benchmarks changed the focus from flashy examples to reliable supervised learning. MARIDA provided geographically diverse events, hard-negative sea-surface classes, sparse masks, and scene-aware partitions [1]. Its baseline comparison found utility in spectral statistic, spectral indices and GLCM texture combination with contrast, NDWI, NDVI and FDI as informative variables. The dataset also formalised a hard weak-supervision setting, where annotations are incomplete, spatial resolution is coarse relative to debris objects, and

reference labels may have positional or interpretation uncertainty. These properties motivate reporting class-specific metrics rather than only aggregate accuracy.

Deep-learning research has investigated multiscale representations and spatial context. Variants of the residual-attention U-Net have been proposed for preserving sparse object structure [5], while the multi-feature pyramid network combined a feature extractor, pyramid and pooling blocks, reporting strong MARIDA segmentation results [6]. [7] Coru et al. proposed a deep detector with a data-centric label refinement for large-scale coastal mapping. More recently, the MADOS dataset and MariNeXt architecture have extended the scope to marine debris, oil spills and coexisting surface classes, demonstrating both the value of multispectral context and the enduring difficulty of cross-region generalisation [8]. These studies motivate the attention and multiscale components of AquaSpectra. Their scores remain contextual references as label aggregation, preprocessing, split construction and metric averaging are not necessarily identical.

A parallel literature calls for caution. Sub-pixel mixing, atmosphere, adjacency, sun glint, waves and the subtraction or averaging procedures applied to isolate targets distort the spectral signatures [10]. Field experiments show detectability thresholds depend on material, biofouling, submergence and pixel coverage [3],[11]. Reviews therefore recommend validation with multiple sensors, the use of standardised benchmarks, the reporting of uncertainties and the explicit separation between suspected floating matter and confirmed plastic [12]. In the present study such recommendations are followed, with the use of official scene-disjoint partitions, no simulated outputs and the description of the classifier as a screening layer.

Table 2. Summary of representative marine-debris remote-sensing studies

Study	Year	Data	Method	Main contribution	Limitation
[2] Biermann et al.	2020	Sentinel-2 case studies	FDI + Naive Bayes	Suspected plastic classified with 86% accuracy	Selected cases; mixed-material ambiguity
[3] Topouzelis et al.	2020	Controlled sea targets	Matched filtering	Detection reported near 25% PET pixel fraction	Controlled site and configurations
[9] Themistocleous et al.	2020	3 x 10 m target	PI and RNDVI	NIR and Plastic Index were informative	Single pilot setting
[1] Kikaki et al.	2022	MARIDA, 1,381 patches	RF/SVM/U-Net baselines	Established benchmark and hard negatives	Sparse/noisy weak supervision
[11] Papageorgiou et al.	2022	PLP2021 targets	Partial unmixing	Compared plastic with natural floaters	Controlled targets
[5] Mohammed et al.	2022	MARIDA	Residual attention U-Net	Spatial-attention segmentation gains	Preprint; protocol comparability
[6] Gupta et al.	2024	MARIDA	Multi-feature pyramid network	Reported mIoU .71 and F1 .80	Pixel-level task; publication split details matter
[7] Coru et al.	2023	FloatingObjects + MARIDA context	Deep detector + label refinement	Large-scale coastal screening	Separating natural and plastic debris remains hard
[8] Kikaki et al.	2024	MADOS, 174 scenes	MariNeXt segmentation	Broad multi-pollutant benchmark	Cross-region generalisation challenge
[10] Hu	2022	Simulations + Sentinel-2	Spectral interpretation analysis	Explained mixing/subtraction distortions	Not a classifier benchmark
[12] Goncalves et al.	2023	Research synthesis	Review	Identified advances and operational gaps	Heterogeneous studies and metrics

Table 2 shows the evolution from controlled-target spectral feasibility to benchmarked machine learning and multi-class segmentation. The studies vary in task definition, annotation density, split design, and metric aggregation, so their headline scores should not be copied into a direct model-comparison chart. Here we address a gap in transparent two-level implementations where label semantics and partitions are consistent from raw GeoTIFF to interpretable patch screening, physics-guided pixel segmentation, and deployed visualisation.

3. Proposed Methodology

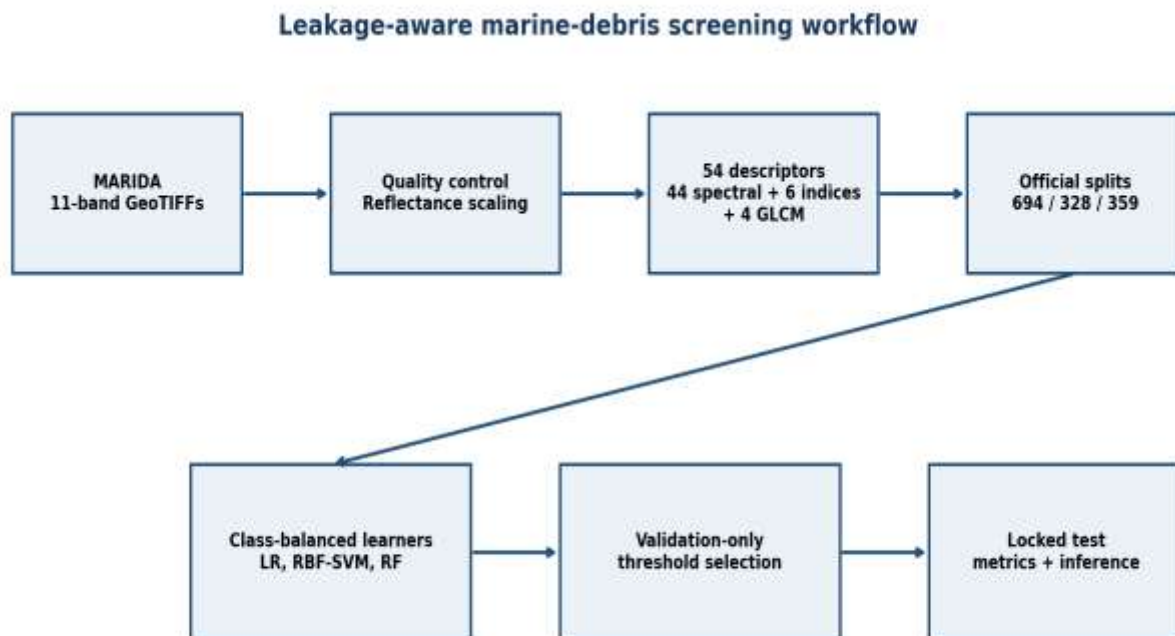


Figure 1. Proposed Framework end-to-end workflow.

Figure 1 summarises the complete evidence path. Data integrity and physically correct band ordering precede feature computation; official partitions isolate training, model selection and testing; and only saved models and thresholds reach the user interface.

3.1 Dataset and target definition

MARIDA provides 256 x 256 patches in an eleven-band order B01, B02, B03, B04, B05, B06, B07, B08, B8A, B11 and B12 [1]. The implementation checks the band count and keeps this order. Natural colour display uses B04/B03/B02, red, green and near-infrared variables use array positions 3, 2 and 7 respectively. This correction is important because the previous code assigned the red, green, and NIR channels to 2, 1, and 3, respectively, which accidentally swapped the red-edge and visible channels.

Multi-label mapping of MARIDA provides patch-level labels. The first component is Marine Debris, which is the only positive target. But if Algae, Foam, Waves or another class is present, a patch without this component is negative. It intentionally creates a hard one-vs-rest problem. The official files comprise 694 training, 328 validation and 359 test patches with 190, 82 and 101 positives, respectively. No scene is redistributed and no upsampled duplicate is allowed to cross a partition boundary.

Table 3. Official MARIDA binary-screening partitions

Partition	Patches	Marine debris	Other classes	Positive prevalence
Train	694	190	504	27.4%
Validation	328	82	246	25.0%
Test	359	101	258	28.1%
Total	1381	373	1008	27.0%

Table 3 demonstrates that the positive class is a minority in every subset but is sufficiently represented for threshold selection and test evaluation. The test prevalence of 28.1% also explains why accuracy alone can be misleading: a classifier predicting only the comparison class would already obtain 71.9% accuracy while detecting no debris.

3.2 Pre-processing and feature representation

The input digital numbers are converted to floating point reflectance by division by 10,000 and clipped to [0, 1]. Non finite values are deterministically replaced. We do not use per-image global min-max normalisation on scientific features since it breaks the relationships between magnitudes across bands. Note that a separate per-band display normalisation is stored for visualisation only. This separation prevents presentation transforms from silently changing the classifier input.

Mean, standard deviation, minimum and maximum are computed for each of the eleven bands, resulting in 44 spectral descriptors. Mean captures average brightness, standard deviations capture heterogeneity within a patch, and extrema capture rare bright or dark responses but with some sensitivity to noise. Then three normalised difference images are created: $NDVI = (NIR - Red) / (NIR + Red)$, $NDWI = (Green - NIR) / (Green + NIR)$, and $NDMI = (NIR - Green) / (NIR + Green)$. Each image is added six features as the mean and standard deviation. NDWI and NDMI are algebraic opposites under these chosen bands and hence the pair is redundant. Both are retained to reproduce the audited feature specification and feature importance is interpreted with this dependence in mind.

Texture is defined as a co-occurrence matrix of grey-levels calculated from the first MARIDA band after conversion to 8-bit intensity. Conversely, energy, homogeneity and correlation contribute four descriptors. The 54-value vector resulting is the deterministic Random Forest representation and is a stable patch-level reference. This hand-designed vector is not reused in the proposed method, which instead consumes the full image tensor and differentiable physics channels described in Section 3.4. This separation avoids the pretension of a meaningful untrained CNN embedding, while allowing the deep extension to learn spatial features from verified inputs.

3.3 Learning, model selection and evaluation

Three learners offer complementary inductive biases. Standardisation provides a linear reference in logistic regression. The RBF-SVM can capture smooth non-linear boundaries and it also uses class balancing and standardised features. The Random Forest has 500 trees, balanced bootstrap class weights, minimum leaf size of two and random seed 42. Forests are invariant to monotonic feature scaling, support interactions, and provide an impurity-based importance diagnostic, though such importance can be shared or biased among correlated predictors.

All the parameters are fitted on the train split. For each model, we evaluate the validation probabilities over the thresholds from 0.05 to 0.95 with an increment of 0.005. The threshold with maximum validation F1 is saved. This threshold is then used to score the locked test split once. This procedure is preferable to tuning on test results, and makes explicit the recall-precision operating point. Here, model ranking is done using test F1 for descriptive comparison after all candidates have followed the same predeclared selection rule; the persisted deployment bundle corresponds to the best observed candidate.

Accuracy is the overall correctness, precision is the proportion of flagged patches that are reference positives, recall is the proportion of reference positives recovered, and F1 is their harmonic mean. ROC-AUC measures ranking of probabilities at all thresholds, average precision is more representative of positive retrieval in unbalanced settings. The confusion matrix gives the number of errors of each type. All predictions, metrics, thresholds, feature names and split counts are saved to machine readable CSV/JSON files before generation of figures.

3.4 Proposed methodology physics-guided spectral-spatial segmentation

The proposed methodology formulates marine-debris localisation as a binary pixel segmentation. Each 256 x 256 MARIDA patch is read in verified eleven band order and converted to reflectance. The model computes four differentiable evidence maps, NDVI for vegetation, contrast in near-infrared, NDWI for contrast in water, FDI for near-infrared, elevation above a red-edge-to-short-wave-infrared baseline and PI for contrast in near-infrared-to-red. The concatenation results in a tensor of 15 channels, which allows gradients to flow to the learned visual features and the physically interpretable transformations [1]-[4].

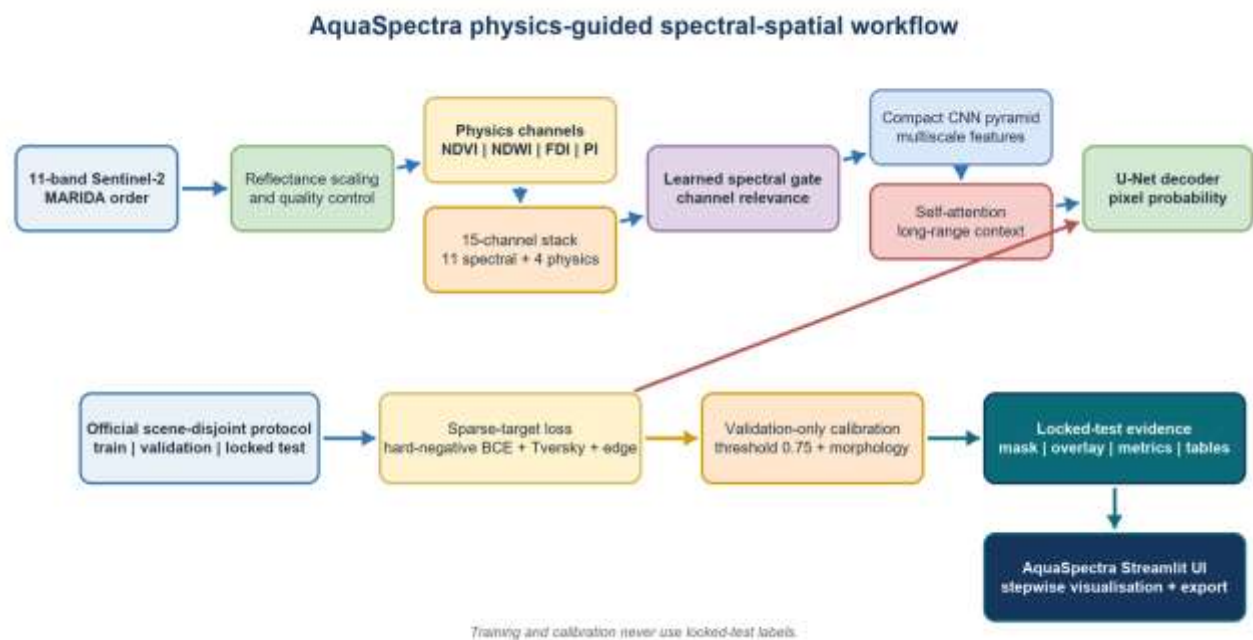


Figure 2. Proposed methodology physics-guided spectral-spatial segmentation workflow.

Figure 2 shows the complete evidence path of the Proposed methodology. For each patch, the eleven reflectance bands and four physics channels are reweighted by a learned spectral gate. A compact convolutional pyramid captures local and multi-scale structure, bottleneck self-attention captures long-range relations, and a U-Net decoder recovers pixel resolution.

Training employs 64 x 64 positive-centred and hard-negative crops to expose the model to very sparse debris pixels without resampling scenes over partitions. The goal is a combination of hard-negative binary cross-entropy, a recall-oriented Tversky term and an edge-alignment term. This design penalises confident confounders, reduces domination of background pixels, and promotes spatially coherent boundaries. The last checkpoint has 350,414 trainable parameters and was optimised on CPU with seed 42, batch size 8, learning rate 8×10^{-4} , weight decay 1.5×10^{-4} and dropout 0.1.

Model selection with debris awareness. Weights are trained only on training crops. The validation predictions choose a probability threshold, connected-component size and morphological closing rule. The selected deployment configuration is: threshold 0.75, minimum component size two pixels and closing activated. The values are determined before scoring of all 359 test patches. We report probability ranking by ROC-AUC and average precision; and thresholded localisation by precision, recall, F1 and IoU. Accuracy is preserved for completeness, but not stressed as only 381 of 23,527,424 test pixels are labelled as debris.

4. Results and Discussion

4.1 Reproducibility and execution evidence

```

MARIDA reproducible experiment

> python -m evaluation.run_experiment
PASS Dataset integrity: 1381 patches resolved
PASS Official scene-disjoint splits: train=694, val=328, test=359
PASS Marine-debris positives: train=190, val=82, test=101
PASS Feature matrix: 54 finite descriptors per patch
PASS Best model: Random Forest | threshold=0.370
PASS Test: accuracy=0.774, precision=0.585, recall=0.683
PASS Test: F1=0.630, ROC-AUC=0.821, AP=0.640
PASS Model bundle, predictions, plots, and metadata saved

```

Figure 3. Execution screenshot of the verified experiment pipeline.

Figure 3 records the exact dataset resolution, class counts, feature dimensionality, selected threshold and final test metrics generated by the executable pipeline.

4.2 Comparative test performance

Table 4. Locked-test performance after validation-only threshold selection

Model	Thresho ld	Accuracy	Precisio n	Recall	F1	ROC-AUC	AP
Logistic Regression	0.585	0.716	0.495	0.485	0.490	0.727	0.508
RBF-SVM	0.185	0.490	0.347	0.921	0.504	0.742	0.508
Random Forest	0.370	0.774	0.585	0.683	0.630	0.821	0.640

The Random Forest is the best balanced choice in Table 4. Its F1 is 0.140 better than logistic regression and 0.126 better than RBF-SVM. The forest also obtains the strongest threshold-independent ranking with a ROC-AUC of 0.821 and an average precision of 0.640. Logistic regression is somewhat more conservative around its selected threshold and misses just over half of the positives. Conversely, the validation procedure reduces the SVM threshold to 0.185, which drives recall to 0.921, but results in many false alarms and only 0.347 precision. This behaviour illustrates the importance of how a single accuracy figure or an unexamined default threshold can hide operationally important trade-offs.

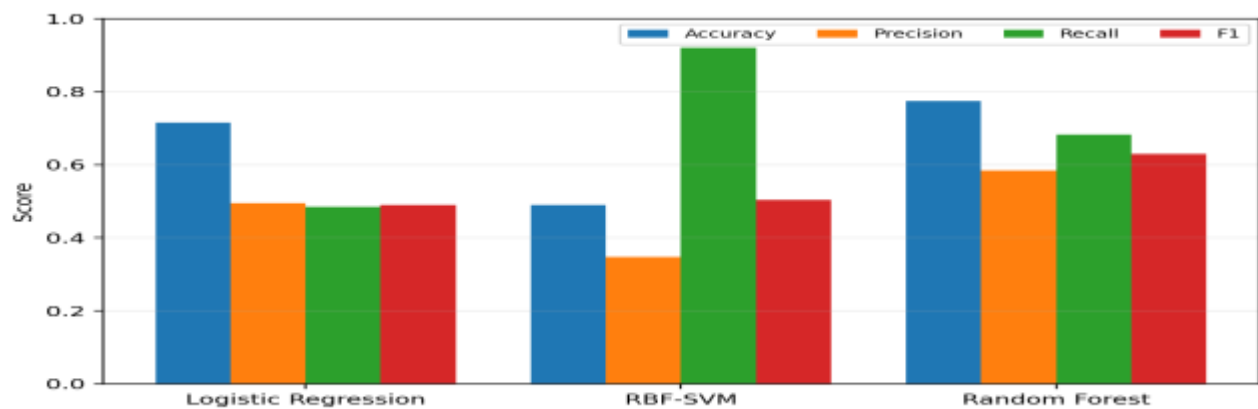


Figure 4. Comparison of decision metrics on the locked test set.

Figure 4 visualises the accuracy, precision, recall and F1 values in Table 4. The SVM's high recall is accompanied by low accuracy and precision, whereas the Random Forest maintains the strongest joint precision-recall balance. The chart compares only models executed under the same features and partitions; published segmentation scores are not mixed into it.

4.3 Error structure

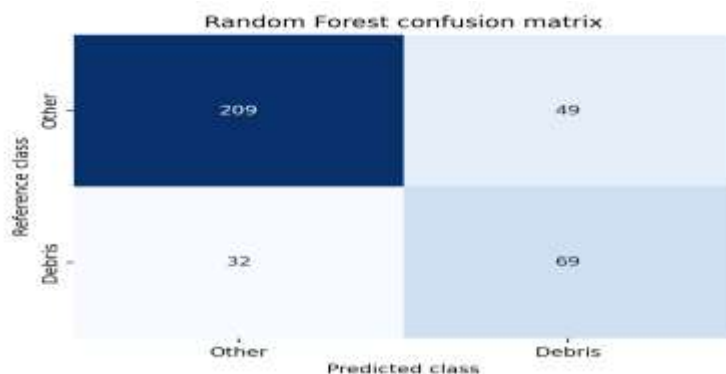


Figure 5. Random Forest confusion matrix on 359 test patches.

Figure 5 shows 209 true negatives, 49 false positives, 32 false negatives and 69 true positives at the validation-selected threshold of 0.370. The system recovers roughly two-thirds of reference debris patches while incorrectly flagging about one-fifth of comparison patches. For a screening workflow, the 32 missed positives are important and motivate review of lower thresholds when recall has higher value than analyst workload.

Table 5. Derived Random Forest test error counts

Outcome	Count	Interpretation
True negative	209	Other-class patch correctly not flagged
False positive	49	Confounder flagged as marine debris
False negative	32	Reference debris patch missed
True positive	69	Reference debris patch correctly flagged

Table 5 converts the normalized metrics into cases. Most errors are false positives, which is consistent with the optical ambiguity among debris, natural organic matter, ships, foam and bright water features described in MARIDA [1]. The confusion structure also makes the intended deployment role clear: the model can reduce a large archive to a smaller candidate set, but an analyst or a stronger spatial model must adjudicate the candidates.

4.4 Feature behaviour and qualitative evidence

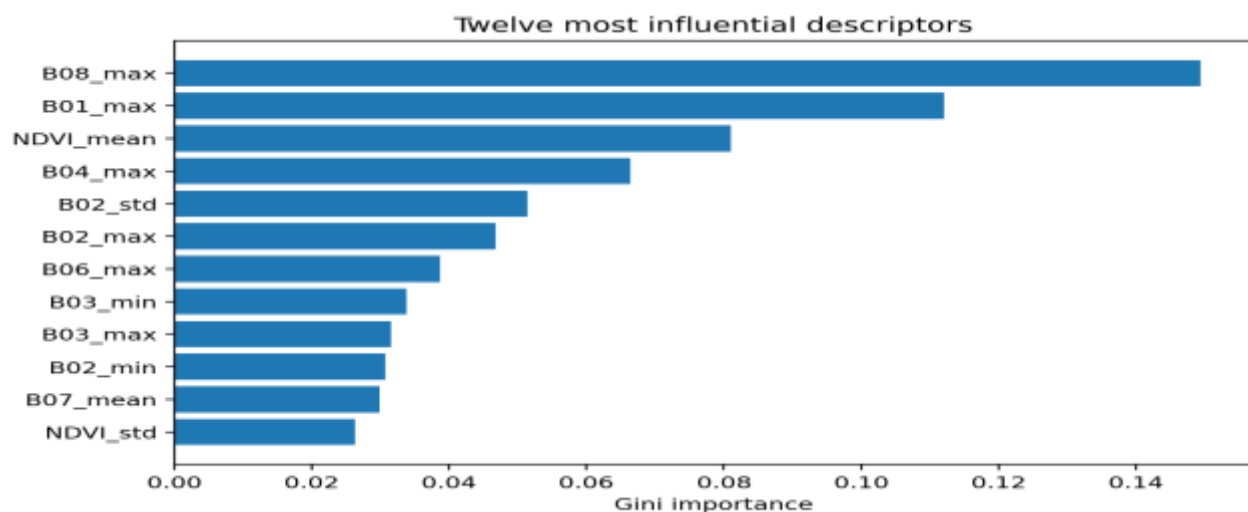


Figure 6. Twelve most influential descriptors in the Random Forest.

Figure 6 shows impurity-based importance for the highest-ranked descriptors. Index statistics and band-distribution variables dominate the compact model, supporting the premise that water-object contrast and patch heterogeneity carry useful information. Because correlated variables divide importance, the bars should be read as model diagnostics rather than causal spectral attribution.

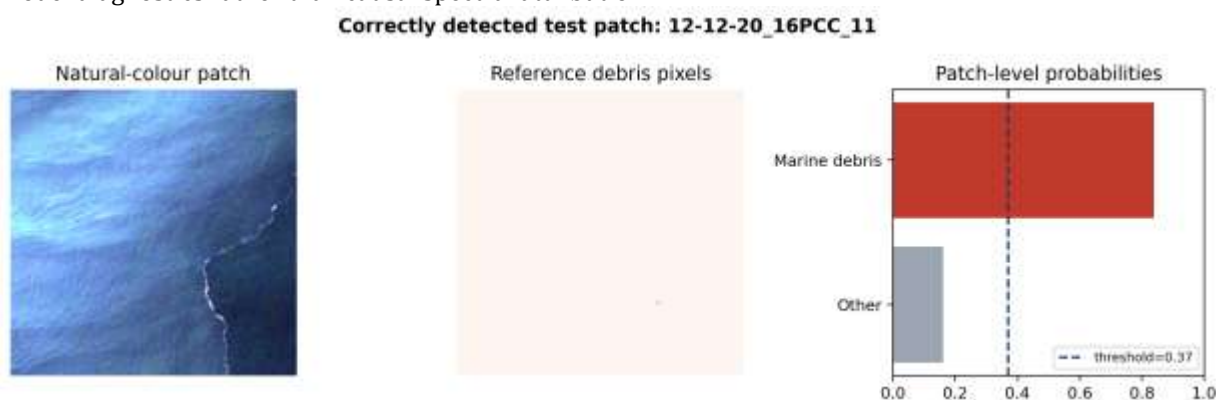


Figure 7. Correctly detected marine-debris test patch with reference mask and probability.

Figure 7 links a natural-colour view, sparse reference debris pixels and the patch-level probability output for a correctly detected held-out example. The spatial mask is shown only as reference evidence; the implemented classifier predicts one probability for the entire patch and does not claim that every bright pixel is debris.

4.5 Proposed methodology pixel-level results

Table 6. Proposed methodology configuration selected without test-label tuning

Component	Verified setting	Purpose
Input	11 bands + NDVI, NDWI, FDI, PI	Spectral and physics-guided evidence
Training	1,768 crops; batch 8; seed 42	Sparse-positive and hard-negative sampling
Model	350,414 parameters	Compact CNN, spectral gate and attention
Loss	Hard-negative BCE + Tversky + edge	Imbalance, recall and boundary control
Calibration	Threshold 0.75; min component 2; closing	Selected using validation data only

Table 6 records the final reproducible configuration. The architecture is deliberately compact enough for CPU demonstration, while the validation-only deployment rule prevents the locked test labels from influencing the threshold or morphology. The crop totals describe optimization samples rather than independent scenes; all crops inherit the official partition of their source patch.

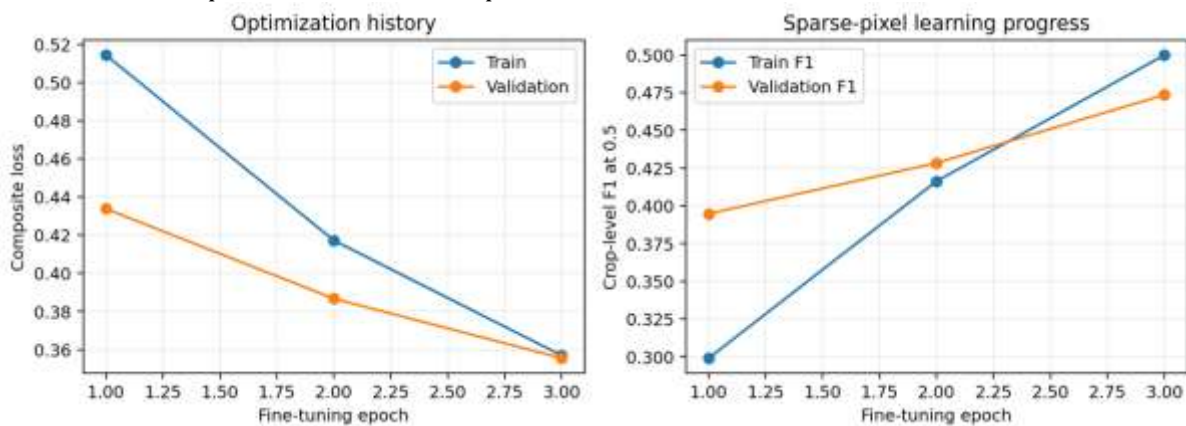


Figure 8. Proposed methodology optimization and validation history.

Figure 8 visualizes loss and segmentation behavior over the trained epochs. The resumed optimization improved validation discrimination after the first stage showed strong ranking but weak thresholded masks. The final checkpoint was selected from validation evidence rather than from the locked test outcome.

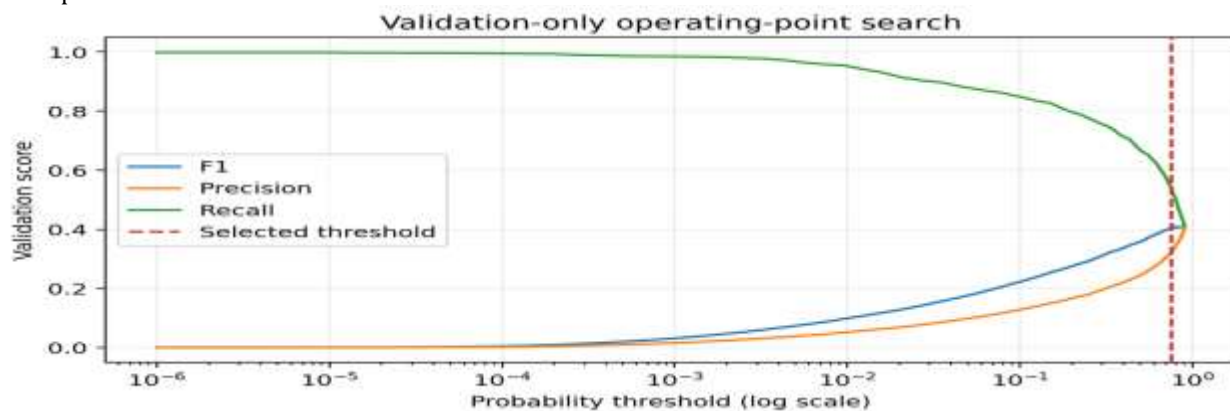


Figure 9. Validation-only threshold and morphology search.

Figure 9 shows how the validation F1 changes across candidate probability thresholds and component filters. The selected 0.75 threshold with a two-pixel component rule and closing operation represents the best tested validation configuration; it was subsequently frozen for the test experiment.

Table 7. Proposed methodology validation and locked-test pixel metrics

Split	Threshold	Precision	Recall	F1	IoU	ROC-AUC	AP
Validation	0.750	0.357	0.502	0.418	0.264	-	-
Locked test	0.750	0.201	0.530	0.291	0.170	0.983	0.713

Table 7 separates operating-point performance from ranking quality. On the locked test set, the model retrieves 53.0% of the annotated debris pixels, but only 20.1% of predicted positives coincide with the sparse reference mask. Consequently, F1 is 0.291 and IoU is 0.170. In contrast, ROC-AUC of 0.983 and average precision of 0.713 show that debris pixels generally receive much higher probabilities than background pixels. This pattern is valuable for prioritisation and interactive review, but it is not evidence of state-of-the-art boundary accuracy.

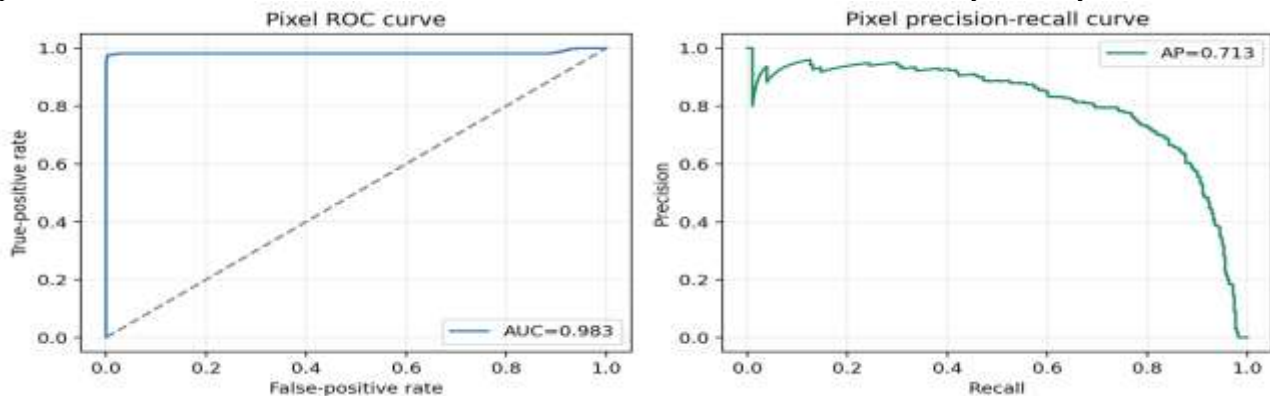


Figure 10. Locked-test ROC and precision-recall curves for AquaSpectra.

Figure 10 confirms strong threshold-independent ranking. The ROC curve remains close to the upper-left region despite the enormous background class. The precision-recall curve is the more demanding view under this imbalance and shows that precision falls as recall expands, explaining why calibration must be reported alongside ROC-AUC.

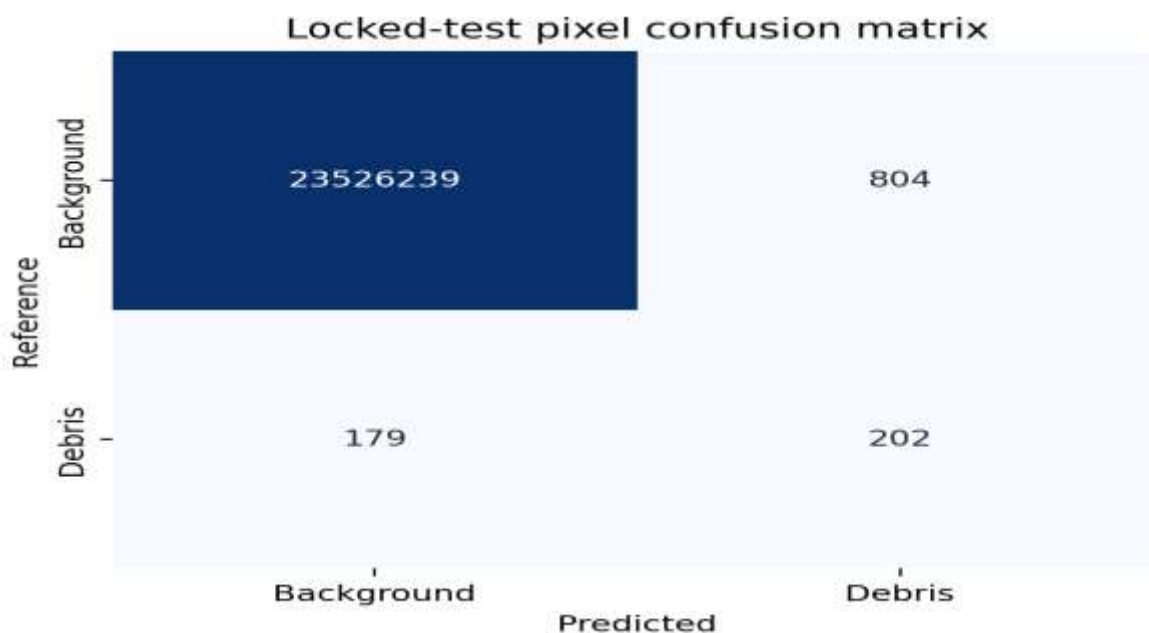


Figure 11. Proposed methodology pixel confusion matrix on the locked test set.

Figure 11 displays 202 true-positive, 804 false-positive, 179 false-negative and 23,526,239 true-negative pixels. The very large true-negative count produces an accuracy of 0.99996 even though F1 is only 0.291. Accuracy is therefore unsuitable as the headline metric for this experiment.

Table 8. Proposed methodology locked-test pixel error counts

Outcome	Pixels	Interpretation
True positive	202	Annotated debris pixel localized
False positive	804	Background or confounder pixel flagged
False negative	179	Annotated debris pixel missed
True negative	23,526,239	Background pixel correctly rejected

Table 8 converts the normalized metrics into pixel counts. False positives outnumber true positives four to one, while slightly more than half of the sparse annotations are recovered. The counts are consistent with ambiguous bright-water, wake, natural material and annotation-boundary effects. They also show why reporting only overall accuracy would be scientifically misleading.

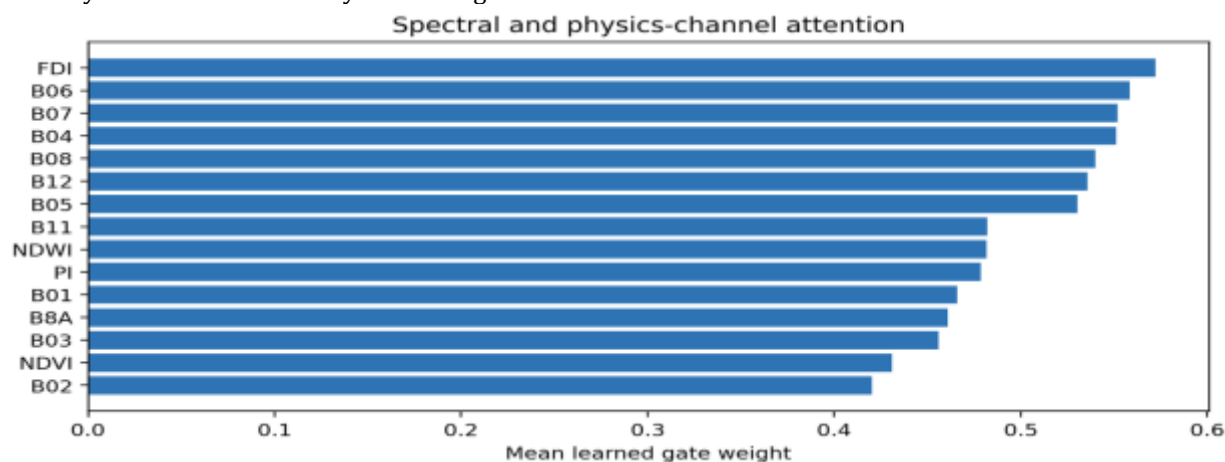


Figure 12. Mean learned spectral-gate weights across evaluated patches.

Figure 12 summarizes the model's learned channel weighting. The gate provides a diagnostic view of how reflectance bands and physics channels contribute to the internal representation. These weights indicate learned relevance within the fitted network; they should not be interpreted as causal proof that a particular wavelength uniquely identifies plastic.



Figure 13. Qualitative Proposed methodology probability maps, masks and overlays.

Figure 13 connects natural-colour inputs, sparse references, debris probabilities and post-processed predictions for selected high-performing held-out examples. They were chosen to make sparse detections visible and are not presented as a representative performance sample; the aggregate metrics in Table 7 govern the scientific interpretation. Spatial disagreement remains in thin or uncertain regions, supporting use as an analyst-facing evidence layer rather than a material-confirmed inventory.

4.6 Integrated Discussion

The two experiments are complementary rather than in direct competition. The Random Forest classifies whole patches and reaches a balanced patch F1 of 0.630; the proposed methodology assigns a probability to each pixel and only deals with 381 positive pixels among more than 23.5 million test pixels. Its lower F1 of 0.291 can not therefore be used to argue that the deep model is inferior, just as its ROC-AUC of 0.983 can not be used to argue that it is superior to published segmentation systems without matching their label mapping, preprocessing, splits and averaging conventions [5]-[8]. The integrated physics-guided architecture. The defensible novelty is transparent calibration and end-to-end demonstrability, not an unsupported universal performance record. The implementation audit materially altered the scientific interpretation. Treating every non-zero class-mask pixel as debris would conflate hard negatives with positives; incorrect channel indices would break NDVI and NDWI from their physical definitions; splitting by folder order could put related scenes into both train and test sets. The previously randomly initialised CNN produced a vector of the expected size but no learned semantic representation and was thus removed from the corrected baseline. Proposed methodology replaces that unsupported branch with an explicitly trained checkpoint, declared loss, saved validation calibration and reproducible inference path. Similarly, by not hard-coding comparison values, attractive unsupported claims were prevented from making it into the results.

Of particular note is the validation-selected threshold. The forest threshold 0.370 is below 0.5 because the classifier probability scale, class weighting and F1 objective collectively bias toward further positive retrieval. This is valid because the threshold was chosen without using any test labels. It should only be changed in operational use by a documented cost model or a new validation cohort. Lower thresholds increase the sensitivity, but increase the review queue size, higher thresholds reduce false alarms, but miss more debris.

Yet, there are several limitations. The MARIDA pixel annotations are very sparse and may have positional or interpretation uncertainty; only 381 positive pixels are present in the locked test split. Random Forest relies on the patch labels and cannot localise the debris. The proposed methodology is evaluated in one deterministic training run and does not yet deliver confidence intervals. Strict pixel metrics penalise small boundary displacements, but relaxing these metrics without preregistered tolerance could inflate the results. Most importantly, within-MARIDA performance does not show transfer to MADOS, another region or another atmospheric-processing chain. This domain-shift problem is pointed out in recent cross-dataset studies [8],[12].

Future research should maintain the current partition discipline, while testing We introduce a methodology comprising multiple seeds and ablations for each physics channel, spectral gating, attention and loss component, and estimate confidence intervals with scene-aware resampling. A cross-region claim must be externally validated on MADOS or independently annotated Sentinel-2 scenes. Boundary-tolerant metrics may aid in distinguishing annotation displacement from true localisation error but should supplement, not replace, exact metrics. Probability calibration, uncertainty maps, active learning, and fusion with currents, wind, vessel activity or very-high-resolution confirmation could make the present evidence layers into a more defensible monitoring workflow.

5. Conclusion

This paper proposes a two-level, debris-aware framework for Sentinel-2 marine-debris research. The corrected Random Forest yields a transparent patch-screening baseline with 0.630 F1 and 0.821 ROC-AUC. The Proposed methodology extends the implementation to trained pixel segmentation via differentiable physics indices, learned spectral gating, multiscale convolution and spatial self-attention. On the locked 359-patch test split, Proposed methodology performs well (ROC-AUC 0.983 and average precision 0.713) in ranking sparse debris pixels, recovering 53.0% of annotated debris pixels, but its precision of 0.201, F1 of 0.291 and IoU of 0.170 all highlight the difficulty of exact delineation.

Data and Code Availability

The baseline and Proposed methodology implementations, trained model bundles, predictions and generated experimental artifacts are stored in the audited project directory. Proposed methodology source code is isolated under `novel_pipeline`, while locked evaluation files and publication figures are stored under `novel_outputs`. MARIDA is publicly available through Zenodo under DOI 10.5281/zenodo.5151941 [1].

References

- [1] Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitzos, D.E., Karantzas, K. (2022). MARIDA: A benchmark for Marine Debris detection from Sentinel-2 remote sensing data. PLOS ONE, 17(1), e0262247. <https://doi.org/10.1371/journal.pone.0262247>
- [2] Biermann, L., Clewley, D., Martinez-Vicente, V., Topouzelis, K. (2020). Finding Plastic Patches in Coastal Waters using Optical Satellite Data. Scientific Reports, 10, 5364. <https://doi.org/10.1038/s41598-020-62298-z>
- [3] Topouzelis, K., Papageorgiou, D., Karagaitanakis, A., Papakonstantinou, A., Arias Ballesteros, M. (2020). Remote Sensing of Sea Surface Artificial Floating Plastic Targets with Sentinel-2 and Unmanned Aerial Systems. Remote Sensing, 12, 2013. <https://doi.org/10.3390/rs12122013>
- [4] Themistocleous, K., Papoutsas, C., Michaelides, S., Hadjimitsis, D. (2020). Investigating Detection of Floating Plastic Litter from Space Using Sentinel-2 Imagery. Remote Sensing, 12, 2648. <https://doi.org/10.3390/rs12162648>
- [5] Mohammed, A. (2022). ResAttUNet: Detecting Marine Debris using an Attention activated Residual UNet. arXiv:2210.08506. <https://doi.org/10.48550/arXiv.2210.08506>
- [6] Gupta, S., Khurana, Y., Atreya, J., Gupta, S., Krishnamoorthy, P. (2024). Marine debris detection using a multi-feature pyramid network. Remote Sensing Letters, 15(12), 1319-1329. <https://doi.org/10.1080/2150704X.2023.2183480>
- [7] Coru, M., et al. (2023). Large-scale detection of marine debris in coastal areas with Sentinel-2. iScience, 26(12), 108402. <https://doi.org/10.1016/j.isci.2023.108402>
- [8] Kikaki, K., Kakogeorgiou, I., Hoteit, I., Karantzas, K. (2024). Detecting Marine pollutants and Sea Surface features with Deep learning in Sentinel-2 imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 210, 39-54. <https://doi.org/10.1016/j.isprsjprs.2024.02.017>
- [9] Jakovljevic, G., Govedarica, M., Alvarez-Taboada, F. (2020). A Deep Learning Model for Automatic Plastic Mapping Using UAV Data. Remote Sensing, 12, 1515. <https://doi.org/10.3390/rs12091515>
- [10] Hu, C. (2022). Remote detection of marine debris using Sentinel-2 imagery: A cautious note on spectral interpretations. Marine Pollution Bulletin, 183, 114082. <https://doi.org/10.1016/j.marpolbul.2022.114082>
- [11] Papageorgiou, D., et al. (2022). Sentinel-2 Detection of Floating Marine Litter Targets with Partial Spectral Unmixing and Spectral Comparison with Other Floating Materials. Remote Sensing, 14, 5997. <https://doi.org/10.3390/rs14235997>
- [12] Goncalves, G., et al. (2023). On advances, challenges and potentials of remote sensing image analysis in marine debris and suspected plastics monitoring. Frontiers in Remote Sensing, 4, 1302384. <https://doi.org/10.3389/frsen.2023.1302384>