Sensors and Materials, Vol. 38, No. 6 (2026) 3421–3440
MYU Tokyo

S & M 4511

Land–Water Discrimination Based on Single-wavelength Waveform Features in Airborne Bathymetric LiDAR

Hyejin Kim[1] and Jaebin Lee[2*]

[1]Social Eco-Tech Institute, Konkuk University
Gwangjin-gu, Seoul 05029, South Korea
[2]Department of Architectural, Civil and Environmental Engineering,
Coast and River Spatial Information Research Lab., Mokpo National University,
Cheonggye-myeon, Muan-gun, Jeonnam 58554, South Korea

(Received December 15, 2025; accepted May 27, 2026)

Keywords: airborne bathymetric light detection and ranging, land–water discrimination, full waveform, machine learning classification, feature selection

Airborne bathymetric Light Detection and Ranging (LiDAR) systems have attracted attention as efficient surveying tools that acquire high-resolution and high-precision coastal topographic data more cost-effectively than traditional shipborne acoustic sounding or field surveys. Since laser pulses are refracted at the air–water interface, for the accurate registration of seafloor points, it is crucial to distinguish whether each return signal is from land or water at the waveform stage, before generating the point cloud. Conventional land–water discrimination techniques often rely on near-infrared (NIR) channel data for water-surface detection or water-body identification. However, NIR signal reliability is often compromised by specular reflection from water surfaces, and many recently developed sensors employ only a single green laser wavelength owing to system miniaturization and weight reduction. This situation underscores the necessity for land–water discrimination techniques that use only single green-channel waveform information. In this study, we analyzed various waveform features extracted from individual waveforms acquired with the Seahawk airborne bathymetric LiDAR system across coastal areas with varying water depths and turbidities. These waveforms were decomposed into Gaussian components, from which features were extracted and used in machine learning classifiers to evaluate their versatility and effectiveness for land–water discrimination under diverse coastal conditions. Four tree-based machine learning models—decision tree, random forest, XGBoost, and LightGBM—were evaluated using a stratified cross-validation scheme for performance assessment. All models achieved a high validation accuracy of approximately 0.99, demonstrating discriminative capability based on waveform features. In comparative evaluations considering both test accuracy and computational efficiency, LightGBM showed the most balanced performance, indicating its suitability as a general-purpose model for waveform-based land–water discrimination.

*Corresponding author: e-mail: lee2009@mokpo.ac.kr
https://doi.org/10.18494/SAM6196

ISSN 0914-4935
https://myukk.org/

1. Introduction

Airborne bathymetric Light Detection and Ranging (LiDAR) systems provide an opportunity to capture both terrestrial and submerged features in a single acquisition.[1,2] An accurate delineation of land–water boundaries is essential for coastal zone management, hydrological modeling, and environmental monitoring. However, discriminating land from water in airborne bathymetric LiDAR (ABL) datasets remains challenging because of mixed returns in shallow waters, variable surface conditions, and the presence of transitional zones such as wetlands and tidal flats.[3,4]

Traditional land–water discrimination approaches rely primarily on discrete-return LiDAR data and intensity-based thresholding. The most straightforward method exploits the differential penetration characteristics of near-infrared (NIR) and green wavelengths: NIR pulses are strongly absorbed by water and produce returns only from land or the water surface, whereas green pulses penetrate the water column and yield bathymetric information.[5,6] By determining the presence or absence of NIR returns, or by analyzing intensity ratios between the two channels, researchers have successfully mapped shorelines and intertidal zones.[7,8] For instance, Guenther *et al.*[9] established baseline protocols for the Scanning Hydrographic Operational Airborne Lidar Survey (SHOALS) system, using intensity thresholds and return-count criteria to distinguish land from water. Similarly, Pe'eri *et al.*[10] developed a decision tree (DT) classifier that integrates NIR intensity, return density, and local surface roughness to automate shoreline extraction. Other studies have incorporated ancillary spatial information, such as elevation continuity and point density variations, to refine classification boundaries.[11,12] While computationally efficient and operationally proven, these discrete-return methods are inherently limited by their reliance on predefined thresholds, which may not generalize across different environmental conditions, sensor configurations, or tidal stages.[13]

To overcome these limitations, recent research has focused on full-waveform LiDAR analysis, which preserves the complete temporal profile of the returned signal and enables a more nuanced discrimination of surface types. Full-waveform data provide rich information about the vertical structure and scattering properties of targets, allowing for the extraction of features such as pulse width, rise time, amplitude, and echo shape asymmetry.[14,15] These features are particularly valuable in transitional zones where discrete returns alone may be ambiguous. For example, Mandlburger *et al.*[16] demonstrated that waveform-derived metrics—including the number of peaks, pulse width variation, and backscatter intensity—could effectively separate land, water surface, and submerged vegetation in shallow coastal environments. Their approach leveraged the characteristic differences in waveform shape: land surfaces typically produce sharp, high-amplitude peaks, whereas water surfaces yield broader, lower-amplitude returns owing to specular reflection and surface roughness.[17] Moreover, the presence of a secondary peak corresponding to the seabed, along with intervening water-column backscatter, serves as a diagnostic feature for water classification.[18,19]

Building on these principles, advanced waveform-based classification frameworks have been proposed in several studies. Allouis *et al.*[20] used wavelet decomposition to analyze the spectral content of waveforms and distinguish water from land on the basis of signal complexity. Pan *et al.*[21] applied continuous wavelet transformation to enhance weak bottom returns and improve land–water discrimination in turbid rivers. More recently, Richter *et al.*[22] have introduced a waveform shape descriptor that quantifies the degree of asymmetry and skewness, demonstrating improved classification accuracy in optically complex waters. Wu *et al.*[23] further advanced this approach by integrating waveform curvature analysis with multichannel fusion, achieving robust land–water separation even under challenging illumination and turbidity conditions. Additionally, machine learning techniques have been incorporated into full-waveform classification workflows. For instance, Mader *et al.*[24] employed random forest (RF) classifiers trained on waveform-derived features to automate land–water boundary detection, achieving accuracies exceeding 95% in diverse coastal settings.

Despite these advances, most existing land–water discrimination methods rely on multi-wavelength LiDAR data, particularly the combination of NIR and green channels, or on discrete-return representations. Such approaches are inherently limited when applied to single-wavelength ABL systems, where only green-channel waveform data are available. Furthermore, few studies have systematically investigated the discriminative capability of full-waveform features alone under varying coastal conditions. Therefore, a robust waveform-based framework that operates independently of multichannel data remains insufficiently explored.

To address this gap, in this study, we propose a waveform-based land–water discrimination framework using single-wavelength ABL data. A set of physically interpretable waveform features is extracted and evaluated using multiple tree-based machine learning models, including DT, RF, XGBoost, and LightGBM. In addition, a feature selection strategy is employed to identify a compact yet effective subset of predictors. The proposed approach is validated across multiple coastal environments to assess both classification performance and generalization capability, with particular attention on computational efficiency and practical applicability.

2. Methods

We develop a supervised framework that (i) decomposes ABL full waveforms into Gaussian components, (ii) extracts physically meaningful waveform features at both component and waveform levels, (iii) learns land–water discrimination with tree-based classifiers, and (iv) selects the optimal feature subset and classifier via wrapper selection (Fig. 1).

2.1 Waveform decomposition

Because the transmitted laser pulse is well approximated by a Gaussian and a superposition of returns from the water surface, bottom, and water-volume backscattering, the returned waveform can be modeled as a Gaussian mixture:

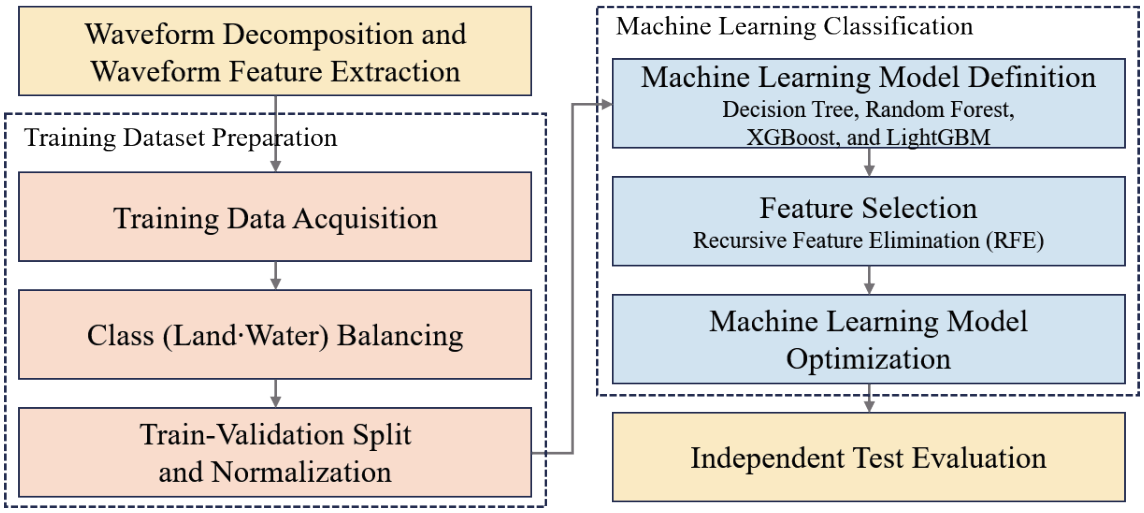


Fig. 1 (Color online) Waveform-based land–water discrimination process.

$$f(t) = \sum_{i=1}^{n} A_i \exp\left(-\frac{(t-\mu_i)^2}{2\sigma_i^2} \right), \tag{1}$$

where n, A_i, μ_i and σ denote the number of Gaussian models, amplitude, temporal position, and standard deviation of the i-th Gaussian component, respectively. ABL waveforms often exhibit a long trailing tail driven by volume backscattering and multiple weak subpeaks caused by suspended matter, bubbles, and small objects, yielding asymmetric, non-Gaussian shapes.[25] Local peaks are identified from the original waveform and used as initialization points for multi-Gaussian fitting.[18] To better handle asymmetric, multi-return structures, we adopt adaptive progressive Gaussian decomposition (APGD).[26,27] This approach improves the waveform decomposition performance by iteratively adding the estimated potential peak candidates to the initial peaks. APGD can effectively decompose irregular ABL waveforms, regardless of water depth or turbidity, and extract various features for each component by representing them as Gaussian functions (Fig. 2).

2.2 Waveform feature extraction

Since each return component is decomposed into a Gaussian model, various features that reflect its physical properties can be derived for each component. Commonly used basic features include amplitude (A), width (W, expressed as full width at half maximum, FWHM), center (C), the number of return (N), and the total number of returns (N_T). Various derivative features have been developed from these basic parameters and applied to tasks such as land–water discrimination[28] and water-layer labeling.[29] In this study, in addition to these component-based features, we extracted waveform-level features that comprehensively characterize each full waveform (Table 1).

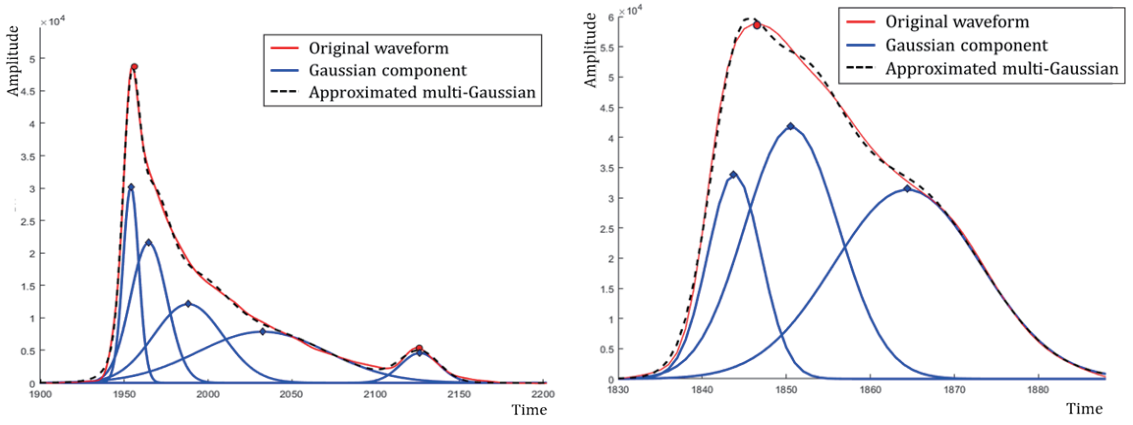
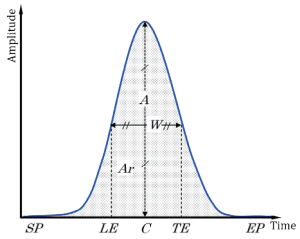
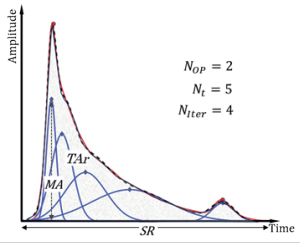


Fig. 2 (Color online) Examples of adaptive progressive Gaussian decomposition: (a) deep and clear water; (b) shallow and turbid water.

Table 1
Definitions of component-level and waveform-level features and their specifications.

Level	Feature	Specification
Component level 	Amplitude	A
	Center	C
	Width (FWHM)	W
	Return number	N
	Leading edge	$LE = C - W/2$
	Trailing edge	$TE = C + W/2$
	Start point	$SP = C - W$
	End point	$EP = C + W$
	Area	$A_r = \sqrt{\pi / (2\ln 2)}AW$
	Area ratio	$R_{Area} = A_r/T_{Area}$
	AW ratio	$R_{AW} = A/W$
	Normalized return	$R_N = N/N_T$
	Center difference	$D_C = C - C_1$
	Leading edge difference	$D_{LE} = LE - LE_1$
	Trailing edge difference	$D_{TE} = TE - TE_1$
Waveform level 	Number of returns	N_T
	Number of original peaks	N_{OP}
	Number of iterations	$N_{Iter} = N_T - N_{OP}$
	Max amplitude	$MAmp = \max A$
	Max area	$MArea = \max A_r$
	Normalized return of max amplitude	$R_{NMAmp} = \mathrm{N}(\mathrm{argmax}A)/N_T$
	Normalized return of max area	$R_{NMArea} = \mathrm{N}(\mathrm{argmax}A_r)/N_T$
	Total area	$T_{Area} = \sum A_r$
	Signal range	$SR = \max EP - \min SP$

- Component-level features. These capture the arrival timing of each return—linked to target elevation—through C, LE, TE, SP, EP; the lags relative to the first return—related to water depth or vertical spacing—through D_c, D_{LE}, D_{TE}; the reflectance/energy of the target or path via A, A_r, R_{Area}; and the echo shape or sharpness via W, R_{AW}.

- Waveform-level features. These summarize the entire echo, describing structural complexity (N_T, N_{OP}, N_{Iter}), total reflected energy (T_{Area}), valid signal range (SR), dominant amplitude/area ($MAmp$, $MArea$), and the normalized return order of the dominant component (R_{NMAmp}, R_{NMArea}).

Because we deliberately exclude any ancillary information beyond the waveform itself (e.g., point geolocation, scan/flight or sensor metadata), we derive and exploit a broad set of physically interpretable features and then identify the subset that is most discriminative for land–water discrimination.

2.3 Training dataset preparation

We compile training data across coastal interfaces that include diverse on-land covers and water bodies with varying depths and turbidity levels. To ensure geographic diversity, we sample from both the west (macrotidal, generally turbid/shallow) and east (microtidal, generally clear/deeper) coasts of the Korean Peninsula, and from multiple estuarine/harbor settings. Collecting data from a wider variety of regions that capture shoreline complexity, bed-slope variability, and differing seasonal regimes can be advantageous. To mitigate class imbalance, we down-sample the majority class to match the minority count. The data were split into training and validation sets at a 7:3 ratio, and features were standardized with Z-score normalization (mean 0, variance 1) for algorithms that may benefit from normalization.[30]

2.4 Machine learning classifiers

Given the nonlinear, heteroscedastic, and noise-prone characteristics of ABL waveform features—along with the presence of correlated predictors and heterogeneous physical scales—we adopt a suite of robust tree-based machine learning models. These algorithms are suited for complex decision boundaries and known to exhibit strong resilience against outliers, non-Gaussian distributions, and multicollinearity.

- DT: The DT classifier is implemented following the classification and regression trees framework, which recursively partitions the feature space on the basis of impurity-reduction criteria such as the Gini index or entropy.[31] DTs offer the advantage of interpretability and require no feature scaling. However, their deterministic structure makes them highly sensitive to small perturbations in the data and prone to overfitting, limiting their generalization capability when used in isolation.
- RF: RF constructs an ensemble of DTs using bootstrap sampling and random feature subspaces.[32] Through bagging and decorrelated model averaging, RF effectively suppresses variance, thereby providing stability under noisy, outlier-contaminated, or structurally complex waveform patterns. Its drawbacks include reduced interpretability compared with a single tree and increased computational and memory costs owing to large ensemble size.
- XGBoost: XGBoost is a gradient boosting framework that iteratively fits new trees to the residuals of previous ones using an additive optimization strategy.[33] With explicit L_1/L_2 regularization, sparsity-aware split finding, and depth-wise tree growth, XGBoost typically

achieves high predictive accuracy even in high-dimensional or imbalanced settings. Nonetheless, this performance often comes at the expense of substantial hyperparameter tuning and increased training complexity.

- LightGBM: LightGBM is an efficient implementation of gradient boosting DT, which employs leaf-wise tree growth with histogram-based splitting, gradient-based one-side sampling, and exclusive feature bundling for high scalability.[34] Developed by Microsoft Research, LightGBM is designed for large-scale machine learning and high-dimensional data environments, offering substantial improvements in training speed and memory efficiency. Its ability to leverage GPU acceleration enables fast training on large waveform datasets. While operationally efficient, the leaf-wise growth strategy may overfit smaller datasets unless regularization is carefully applied, and the resulting model structure can be less interpretable.

All models were executed using CPU-based training with parallel processing enabled. Hyperparameters were selected on the basis of preliminary experimental results to achieve stable performance across different environments. For RF, the number of trees was set to 200 to provide sufficient ensemble diversity. XGBoost and LightGBM were configured with a learning rate of 0.05 and 200 boosting iterations to balance convergence and generalization. The maximum tree depth in XGBoost was limited to six to prevent overfitting, while LightGBM used a leaf-wise structure with 31 leaves. In both boosting models, subsampling (0.9) was applied to improve generalization performance.

Overall, these complementary properties provide a balanced framework for identifying the most discriminative waveform features and modeling complex land–water reflectance patterns inherent to ABL (Table 2).

2.5 Feature selection

To enhance the robustness and efficiency of the DT-based machine learning models applied in this study, we performed feature selection with recursive feature elimination with cross-validation (RFECV).[35] Feature selection was necessary because the full set of waveform-derived features contains substantial redundancy and inter-feature correlation, which can degrade model generalization and inflate variance in tree-based classifiers. RFECV iteratively removes features with the lowest model-based importance and evaluates performance using

Table 2
Classifiers and qualitative properties.

Model	Learning scheme	Key advantage	Overfitting risk	Large-data scalability
DT	Single tree	Highly interpretable	High	Low
RF	Bagging	Stable and robust	Low	Medium
XGBoost	Gradient boosting	High accuracy and precision	Medium-high	High
LightGBM	Gradient boosting	Very fast; strong on sparse/large datasets	Medium	Very high

stratified K-fold cross-validation with the F1-score as the optimization criterion. This procedure automatically identifies the optimal number of features while mitigating overfitting, which is particularly important when dealing with high-dimensional waveform features. Moreover, when combined with tree-ensemble models, RFECV effectively captures nonlinear feature interactions while suppressing irrelevant or weakly contributing variables. By integrating RFECV into our approach, the final feature subset becomes more compact, physically meaningful, and better aligned with the DT mechanisms.

In this study, a stratified K-fold cross-validation scheme with K = 4 was adopted to ensure that class proportions were consistently maintained across all folds during model evaluation. To improve computational efficiency, the feature elimination step size was configured to remove two features per iteration, allowing a balance between processing time and selection granularity.

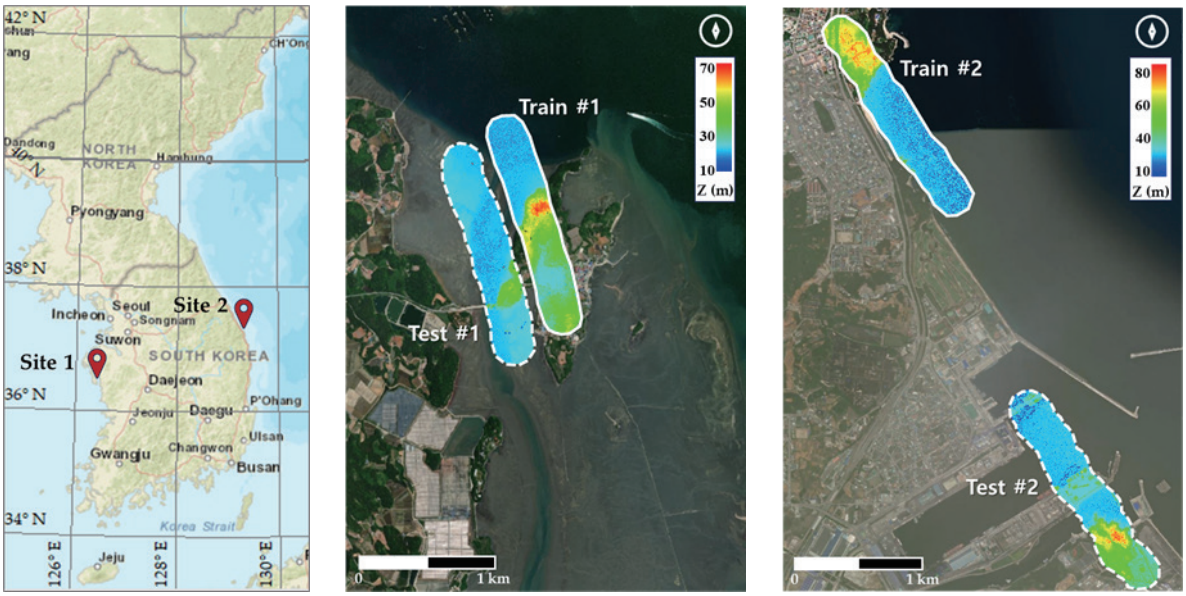
3. Experimental Results

To identify the most effective combination of waveform features and classifiers for land–water discrimination, we trained the models using ABL (Seahawk) waveform data acquired at different times from coastal environments with varying water depths and turbidity levels, and subsequently evaluated their performance on separate test datasets.

3.1 Test datasets

The Seahawk system is a Korean coastal/topographic–bathymetric mapping platform developed by Geostory Inc. with the support of the Ministry of Oceans and Fisheries. It employs a holographic optical element–based circular scanner to acquire co-registered green (532 nm) and NIR (1064 nm) laser returns, and a real-time computation engine that renders 3D point clouds during flight.[36] Since its first operational deployment in July 2018, Seahawk has been extensively used to monitor wetlands and coastal zones in South Korea, enabling simultaneous topographic and bathymetric measurements down to ~35 m water depth (at ~400 m flight altitude) over 2 × 2 m^2 footprint cells in clear water.

The test sites were selected from two coastal regions in Korea that exhibit markedly different environmental characteristics—turbid tidal flats on the west coast and clear-water environments on the east coast—both including land–water boundaries (Fig. 3). The datasets from the two test sites were acquired more than one year apart, and the regions differ substantially in bathymetry, turbidity, and tidal range (Table 3). Site 1 (Hwangdo tidal flat) features shallow waters with a maximum depth of approximately 5 m, gently sloping mud–mixed–sand substrates, a large tidal range (maximum high tide 6.33 m, minimum low tide 2.86 m), and relatively high turbidity [mean suspended particulate matter (SPM) 10.9 mg L^{-1} in 2023, based on the Korean Statistical Information Service].[37] Site 2 (near Mukho Harbor on the east coast) reaches depths of about 11 m and is characterized by a narrow, ridge-shaped coastline, steep bottom slope, small tidal range (approximately 0.25 m), predominantly sandy seabed, and moderate wave-induced coastal erosion, with lower turbidity (mean SPM of 4.6 mg L^{-1} in 2022).[37] For each site, two separate flight strips were acquired. One strip was used as the training dataset, while the other—collected



(a) (b) (c)

Fig. 3 (Color online) Test sites and data overview: (a) locations, (b) test site 1 on the west coast of Korea, and (c) test site 2 on the east coast of Korea. Solid outlines indicate training datasets and dashed outlines indicate test datasets.

Table 3
Test datasets.

	Location	Acquisition date	Maximum depth (m)	Tidal range (m)	Turbidity	Number of waveforms
Site 1	Hwangdo, Taean-gun, Chungcheongnam-do, Korea	15 October 2023	5	2.86–6.33	Medium	Train: 215040 Test: 215040
Site 2	Donghae-si, Gangwon-do, Korea	22 March 2022	11	≈ 0.25	Low	Train: 215040 Test: 215040

along a different flight line—served as the test dataset for independent performance evaluation (Fig. 3).

3.2 Feature selection results

Figure 4 and Table 4 present the importance rankings of the key waveform features selected by each model and the corresponding interpretation. Across all four classification models, a consistent set of waveform-level features emerged as the most influential for land–water discrimination. Specifically, start point, maximum amplitude, total area, signal range, and the normalized return of the maximum-amplitude component were selected as top-ranked predictors. Except for start point, the remaining features are waveform-level descriptors that capture global echo characteristics such as total energy, dominant peak strength, and the temporal extent of valid returns. These properties reflect fundamental physical differences between land and water returns—namely, delayed start times, reduced and attenuated energy

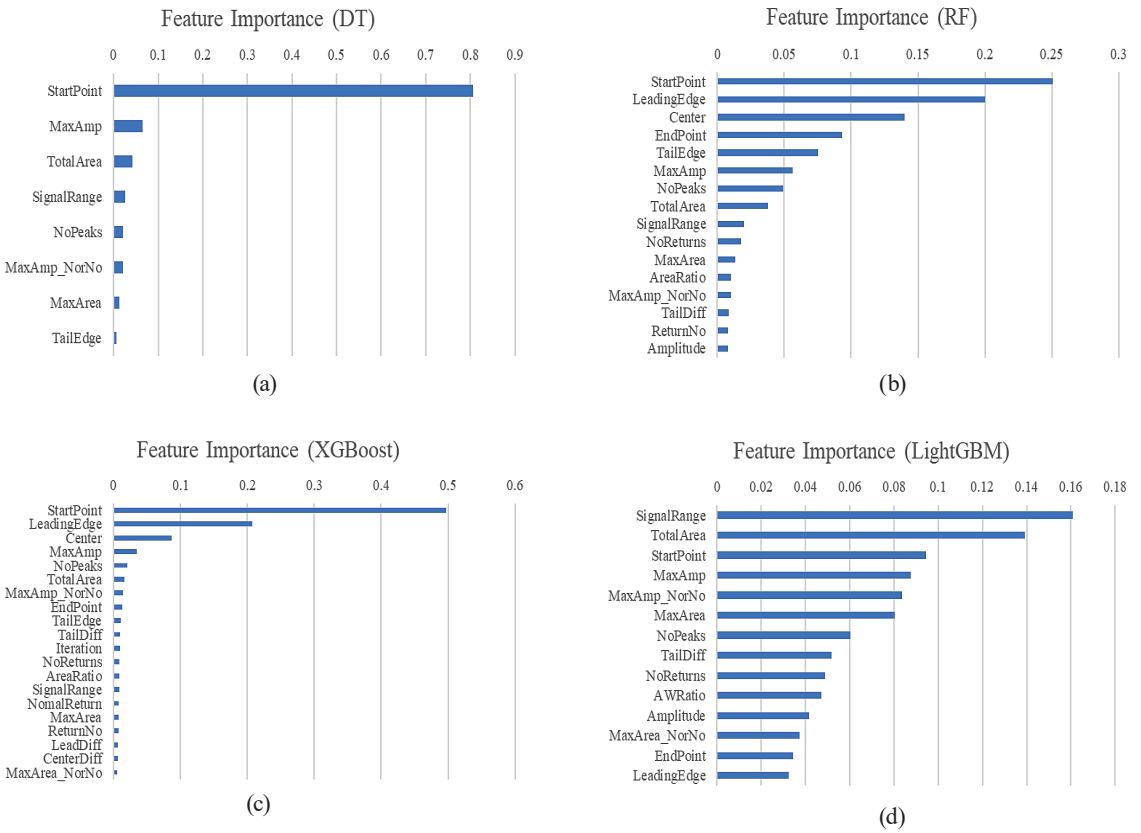


Fig. 4 (Color online) Feature importance results for four classifiers: (a) decision tree (DT), (b) random forest (RF), (c) XGBoost, and (d) LightGBM. Each bar plot shows the relative contribution of waveform features used for land–water discrimination.

Table 4
Selected waveform features for each classifier and their interpretation.

Model	Number of selected features	Major selected features	Interpretation
DT	8	Start point, Max amplitude, Total area, Signal range, Number of original peaks, Normalized return of max amplitude, Max area	· Single-threshold–based features preferred · Immediate class separation using arrival time & max energy · Simple and intuitive decision rules
RF	16	Start point, Leading edge, Center, Max amplitude, Number of original peaks, Total area, End point, Normalized return of max amplitude	· Uses diverse feature combinations across many branches · Captures multiple waveform patterns · Robust to complex and noisy structures
XGBoost	24	Signal range, Total area, Max amplitude, Normalized return of max amplitude, Start point, Max area, Number of original peaks	· Boosting emphasizes the strongest gain-based splits · Arrival-time features repeatedly selected · Produces sharp and highly discriminative boundaries
LightGBM	14	Start point, Leading edge, Center, End point, Trailing edge, Max amplitude, Number of original peaks, Total area	· Leaf-wise tree growth captures global waveform structure · Effectively exploits shape and energy patterns · Highly efficient for large-scale waveform data

due to underwater scattering, and distinct peak-structure complexity—explaining their consistent importance across models.

Model-specific selection patterns also reveal the intrinsic behavior of each learning algorithm. DT and XGBoost place pronounced emphasis on start point, assigning disproportionately high importance to this single feature. This reflects their splitting mechanisms: DT prefers simple threshold-based separations, while XGBoost repeatedly exploits the strongest gain-producing split owing to its boosting nature. As a result, both models tend to rely on arrival-start-time-based discrimination boundaries. In contrast, RF and LightGBM exhibit a more distributed use of features. RF leverages its ensemble structure, drawing on diverse branching rules across multiple trees, which enables the stable modeling of complex waveform patterns such as leading/trailing edges, peak-derived quantities, and energy-related descriptors. LightGBM, because of its leaf-wise growth strategy, frequently selects features that capture global waveform characteristics (e.g., signal range and total area), reflecting its strength in exploiting structural patterns and cumulative energy signatures.

A notable finding is that start point—the moment when the return energy first reaches the sensor—serves as a more discriminative feature than the center of each component. This behavior appears to stem from both the characteristics of the waveform decomposition approach (APGD) used in this study and the fundamental differences between terrestrial and aquatic return signals. For terrestrial surfaces, laser pulses travel through air with negligible attenuation, and reflections typically originate from rigid targets, producing sharply defined peaks. As a result, the temporal gap between start point and center is relatively small. In contrast, as illustrated in Fig. 2, return signals from water environments exhibit markedly different behavior: non-seabed components—such as water-surface reflections, volumetric backscattering from the water column, and even very shallow bottom reflections—tend to begin at nearly the same temporal position when decomposed. Consequently, multiple Gaussian components within a single waveform share similar start-point values, forming a consistent pattern that differs clearly from terrestrial waveforms. This systematic convergence of start points among water-derived components is likely what makes start point a highly discriminative feature for land–water separation.

Overall, the recurring selection of common high-ranked features across all models suggests that these predictors capture physically meaningful differences between land and water returns. Although model behaviors differ, this cross-model agreement indicates that the waveform-based feature set may reflect key aspects of the scattering behavior relevant to land–water discrimination in ABL.

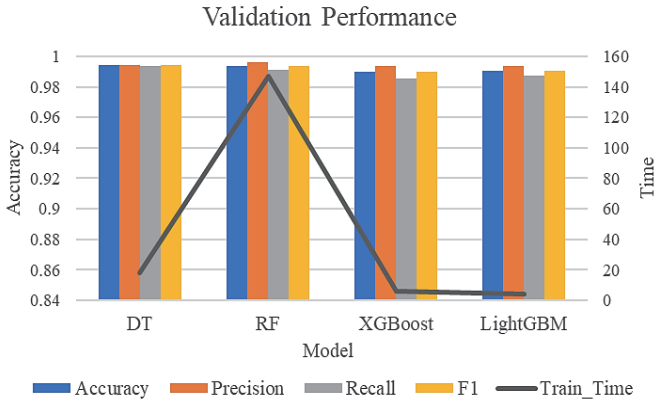
3.3 Classification performance with RFECV

Table 5 and Fig. 5(a) show the validation performance of the four classification models. All models achieved very high overall accuracy (≈0.99), with precision, recall, and F1-scores also exceeding 0.98, indicating that waveform-based land–water discrimination can be learned reliably across diverse tree-based classifiers.

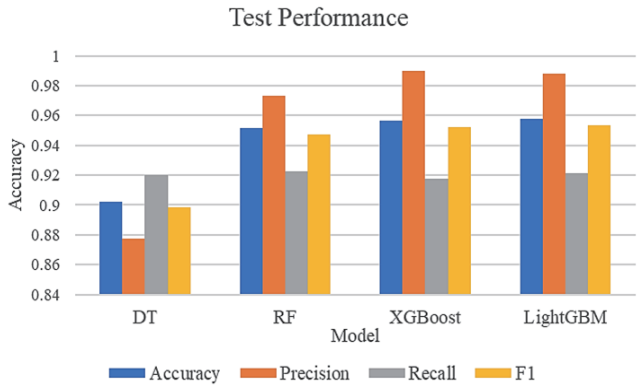
While all four models achieved near-perfect validation performance, the DT classifier exhibited the highest overall validation accuracy (0.9942), slightly outperforming RF, XGBoost, and LightGBM. This result reflects the DT's tendency to exploit a small number of highly discriminative thresholds—particularly those related to strong arrival-time contrasts—that can yield the very sharp separation of land and water samples in the training phase. In contrast, RF showed similarly high scores but required the longest training time by far (processing 430,080

Table 5
Validation accuracy and training time of classification models with RFECV.

Model	Accuracy				Training time (s)
	Overall	Precision	Recall	F1-score	
DT	**0.9942**	0.9945	**0.9939**	**0.9942**	17.77
RF	0.9937	**0.9961**	0.9913	0.9937	146.93
XGBoost	0.9896	0.9935	0.9856	0.9896	6.11
LightGBM	0.9906	0.9938	0.9874	0.9906	**4.02**



(a)



(b)

Fig. 5 (Color online) Performance obtained with RFECV: (a) validation performance (accuracy, precision, recall, F1-score, and training time) for each classifier and (b) test performance evaluated on two test datasets.

waveforms in approximately 146.9 s), owing to the large number of trees and the computational overhead of bootstrap aggregation under the given hardware environment (Intel® Core™ i5-7500 CPU, 16 GB RAM, NVIDIA GeForce RTX 3060 GPU).

LightGBM provided the best balance between performance and computational efficiency. LightGBM was the fastest (processing 430,080 waveforms in approximately 4.0 s) and achieved classification accuracy comparable to that of DT. These results suggest that LightGBM is particularly well suited for large-scale waveform datasets, where rapid training and scalability are key operational requirements. XGBoost also showed fast and excellent training performance, but it fell short of LightGBM's results in both speed and accuracy. Overall, while all models effectively captured the discriminative structure of waveform features, LightGBM offers the most practical performance–efficiency trade-off for large-scale application.

When the trained models were applied to test datasets acquired from different flight lines [Table 6 and Fig. 5(b)], clear differences in generalization performance emerged across classifiers. The RF, XGBoost, and LightGBM models all maintained stable accuracy levels, each achieving overall accuracy and F1-scores close to 0.95, despite the spatial and temporal differences. These results indicate that ensemble-based or boosted DT models are better able to capture waveform-level scattering patterns that generalize across survey conditions.

In contrast, the DT classifier exhibited a marked performance drop on the test set, with overall accuracy and F1-scores decreasing to approximately 0.90. This degradation is consistent with the DT's known sensitivity to overfitting: although it performed well on the validation data, its reliance on a single set of hard thresholds limited its ability to generalize to waveforms collected under different environmental or sensor conditions.

3.4 Classification performance with optimal feature set

To evaluate whether a reduced yet physically meaningful feature set could maintain classification performance, all models were retrained using seven optimal features identified from the feature selection process: start point, total area, maximum amplitude, signal range, the number of original peaks, the normalized return of maximum amplitude, and maximum area.

As shown in Table 7 and Fig. 6(a), the validation accuracy remained above 0.98 for all models, with minimal changes compared with the full feature set. This indicates that the selected features retain the majority of the discriminative information required for land–water discrimination. Test results [Table 8 and Fig. 6(b)] further demonstrate that ensemble and boosting models (RF, XGBoost, and LightGBM) maintain strong generalization performance,

Table 6
Test accuracy of classification models with RFECV.

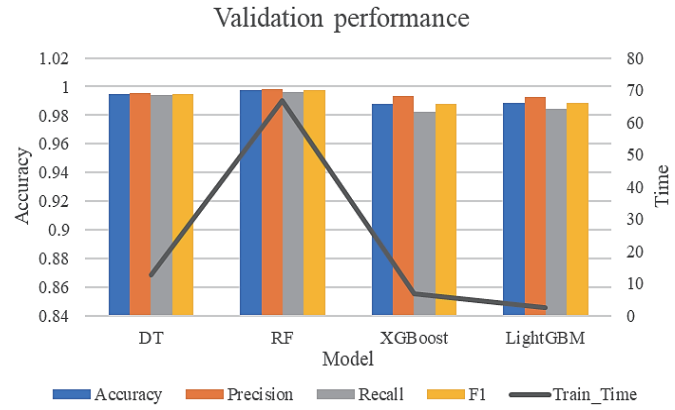
Model	Accuracy			
	Overall	Precision	Recall	F1-score
DT	0.9021	0.8778	0.9202	0.8985
RF	0.9516	0.9732	0.9226	0.9472
XGBoost	0.9565	**0.9896**	0.9173	0.9521
LightGBM	**0.9576**	0.9881	**0.9211**	**0.9534**

with an overall accuracy of approximately 0.95. Among them, LightGBM achieved the highest accuracy and F1-score, confirming its robustness when combined with a reduced feature set.

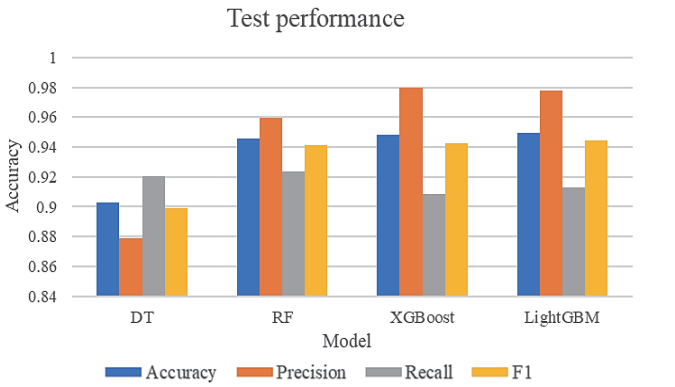
Compared with the RFECV results (Tables 5 and 6), performance differences remained within 1% across both validation and test datasets, suggesting that the excluded features contribute limited additional information. This suggests that the optimal subset retains the core waveform characteristics governing class separability, while removing redundant or weakly

Table 7
Validation accuracy and training time of classification models using the optimal feature set.

Model	Accuracy				Training time (s)
	Overall	Precision	Recall	F1-score	
DT	0.9950	0.9955	0.9944	0.9949	12.64
RF	**0.9975**	**0.9985**	**0.9965**	**0.9975**	66.88
XGBoost	0.9878	0.9931	0.9823	0.9877	6.66
LightGBM	0.9885	0.9928	0.9842	0.9885	**2.44**



(a)



(b)

Fig. 6 (Color online) Performance obtained when using the optimal feature set: (a) validation performance (accuracy, precision, recall, F1-score, and training time) for each classifier and (b) test performance evaluated on two test datasets.

Table 8
Test accuracy of classification models using the optimal feature set.

Model	Accuracy			
	Overall	Precision	Recall	F1-score
DT	0.9028	0.8787	0.9206	0.8991
RF	0.9457	0.9593	**0.9238**	0.9412
XGBoost	0.9480	**0.9796**	0.9084	0.9427
LightGBM	**0.9493**	0.9777	0.9131	**0.9443**

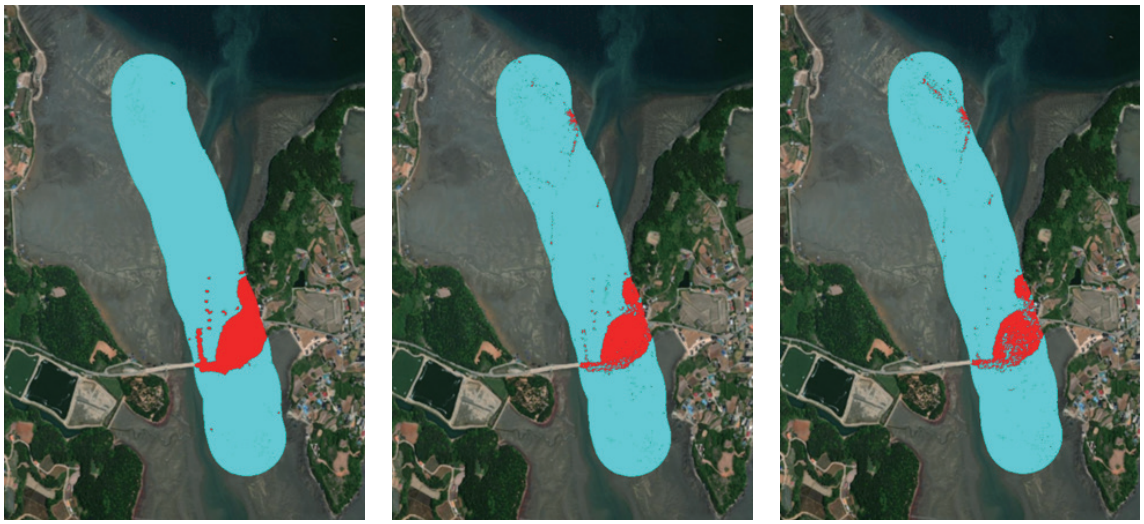
informative parameters that may introduce noise or increase model complexity. The optimal feature subset delivered substantial computational benefits. The training time decreased notably for all models, with LightGBM and XGBoost completing training in only a few seconds. Memory usage and data loading overhead also decreased owing to the smaller number of input dimensions. From an operational standpoint, this improved efficiency is particularly advantageous for large-scale ABL waveform datasets or real-time processing scenarios. Despite these benefits, the reduced feature diversity may limit the model's ability to represent mixed or transitional waveform signatures, particularly in spatially complex environments.

3.5 Visual assessment of land–water discrimination

Figures 7 and 8 present a visual comparison of the LightGBM classification results obtained using the RFECV-selected feature set and the reduced optimal feature subset. The test datasets were intentionally selected from areas containing complex man-made structures where land–water discrimination was particularly challenging. Test #1 corresponds to a region including a major bridge (Hwangdo Bridge) connected to an onshore road, nonfixed mooring facilities, small vessels, and floating bungalows, resulting in a highly heterogeneous land–water interface. The visual inspection of the RFECV-based results indicates that thin docking structures, small boats, floating bungalows, and parts of the bridge deck were occasionally misclassified as water, while some deeper water areas were misclassified as land. Although the classification using the optimal feature subset achieved a comparable quantitative accuracy to the RFECV result, it exhibited a noisier spatial pattern, with increased scattered misclassifications around structural boundaries and mixed regions.

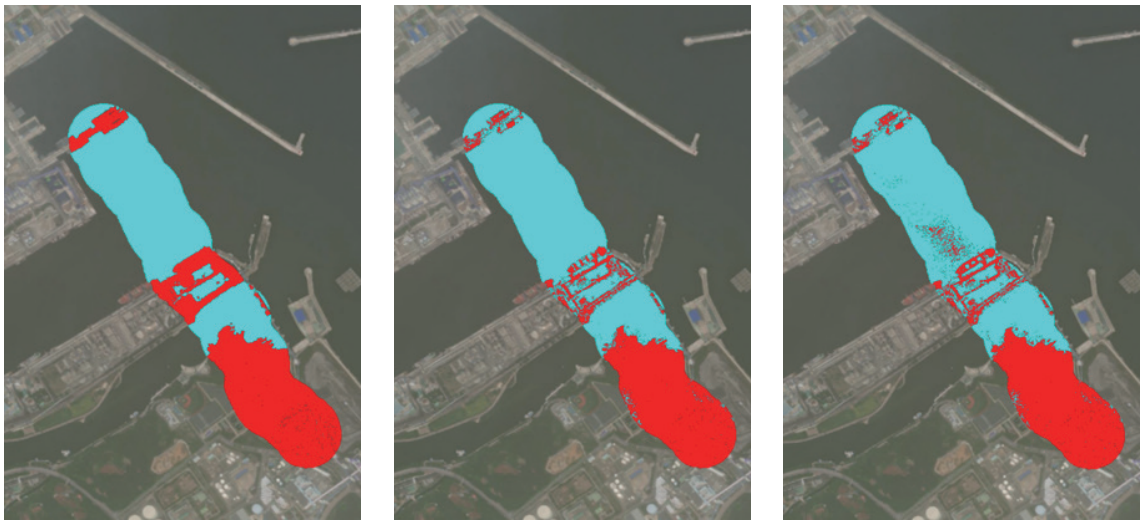
Test #2 represents an industrial harbor environment containing a wet dock filled with water at elevations higher than the surrounding sea level, medium- to large-sized anchored vessels, and cargo-handling piers. In this case, water was overestimated within shallow wet-dock interiors and portions of large ship hulls were also misclassified as water. Similar to Test #1, misclassifications were concentrated around pier structures and the optimal feature subset produced more spatially fragmented and noisy results than the RFECV-based classification.

The frequent occurrence of misclassification near bridges, piers, and vessels can be attributed to the coexistence of land and water surfaces within a single waveform instantaneous field of view (IFOV), leading to mixed or ambiguous waveform signatures. Such mixed returns reduce the separability of land and water classes regardless of the feature selection strategy. While isolated noisy misclassifications within homogeneous land or water regions could be effectively



(a) (b) (c)

Fig. 7 (Color online) Experimental results of Test #1 dataset (cyan: water; red: land): (a) ground truth, (b) classification result with RFECV, and (c) classification result with an optimal feature set.



(a) (b) (c)

Fig. 8 (Color online) Experimental results of Test #2 dataset (cyan: water; red: land): (a) ground truth, (b) classification result with RFECV, and (c) classification result with an optimal feature set.

mitigated through statistical filtering or spatial post-processing, this approach may inadvertently remove small artificial objects such as boats or floating structures. However, since these mobile or small-scale artificial features are not the primary targets of land–water discrimination, their removal is unlikely to significantly affect the overall applicability of the classification results.

Overall, while both approaches achieve comparable quantitative performance, the RFECV-based feature set provides improved spatial consistency in complex environments, highlighting the importance of incorporating a broader range of waveform features when spatial coherence is a critical consideration.

4. Discussion

Although the models achieved high overall accuracy, misclassifications were frequently observed in spatially complex environments such as in areas with bridges, piers, and harbor structures. These areas often produce mixed waveform signatures owing to the coexistence of land and water within a single laser footprint. This limitation highlights the inherent difficulty of waveform-based classification in heterogeneous environments and suggests that additional contextual or spatial information may further improve performance. From a practical perspective, the proposed waveform-based approach is particularly advantageous for modern single-wavelength ABL systems, where auxiliary channels are unavailable. The reduced feature set also enables efficient processing, making the method suitable for large-scale coastal mapping and potential real-time applications.

5. Conclusions

In this study, we investigated land–water discrimination using single-wavelength ABL full-waveform features and tree-based machine learning classifiers. By decomposing raw waveforms through APGD and extracting physically interpretable component-level and waveform-level features, we demonstrated that reliable land–water separation can be achieved without relying on ancillary channels such as NIR or external spatial information. Across diverse coastal environments characterized by different water depths, turbidity levels, and tidal regimes, ensemble and boosting-based classifiers—particularly RF, XGBoost, and LightGBM—exhibited strong generalization capability, maintaining stable accuracy on independent test datasets acquired along separate flight lines.

Feature-selection experiments revealed that a small subset of waveform features—most notably start point, total area, maximum amplitude, signal range, the number of original peaks, the normalized return of the maximum amplitude component, and maximum area—consistently captured the dominant physical differences between land and water returns. Models trained with this reduced optimal feature set achieved validation and test accuracies comparable to those obtained using the full feature set, with only marginal (<1%) differences in performance. This result suggests that the discriminative information for land–water separation is largely concentrated in a limited number of physically meaningful waveform descriptors, and that most secondary features contribute little additional benefit in terms of global accuracy metrics.

However, visual assessment revealed differences between the results of using the RFECV-based feature set and the reduced optimal subset. Although quantitative accuracy metrics were similar, classifications based on the optimal feature subset exhibited more spatially fragmented and noisy patterns, particularly around complex coastal structures such as bridges, piers, docks,

and vessels. These discrepancies are likely attributable to the reduced feature diversity, which limits the model's ability to capture subtle waveform variations associated with mixed land–water returns within a single IFOV. In contrast, the RFECV-based approach—by retaining a broader range of complementary waveform features—produced more spatially coherent classification results in structurally complex environments.

Overall, the results indicate that RFECV-based feature selection provides a more robust balance between quantitative accuracy and spatial consistency, whereas the reduced optimal feature set offers computational efficiency at the cost of increased visual noise. These findings suggest that the choice between feature-selection strategies should be guided by application requirements—favoring RFECV-based models when spatial coherence is critical and compact feature sets when processing speed and scalability are prioritized.

In this study, model training was conducted using datasets acquired from two sites with distinct environmental characteristics and different acquisition periods. Nevertheless, as indicated by the test results, misclassifications tend to occur around coastal structures such as bridges, piers, and vessels, as well as in areas with complex land–water boundaries. This suggests that further exposure to a wider variety of coastal objects and more heterogeneous shoreline configurations during training could lead to a more robust and better-generalized classification model. Therefore, future work will focus on expanding the training and validation datasets to encompass more diverse coastal environments and structural conditions, with the aim of improving model generalization and reliability across broader operational scenarios.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2021R1I1A3059263).

References

1 J. L. Irish and W. J. Lillycrop: ISPRS J. Photogramm. Remote Sens. **54** (1999) 123. https://doi.org/10.1016/S0924-2716(99)00003-9
2 G. C. Guenther: Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd ed., D. Maune (ASPRS Publications, Bethesda, MD, USA, 2007) pp. 253–320.
3 T. Webster, K. McGuigan, N. Crowell, K. Collins, and C. MacDonald: J. Coast. Res. **76** (2016) 31. https://doi.org/10.2112/SI76-004
4 C. E. Parrish, J. A. Dijkstra, J. P. M. O'Neil-Dunne, and C. McKenzie: J. Coast. Res. **76** (2016) 200. https://doi.org/10.2112/SI76-017
5 R. C. Hilldale and D. Raff: Earth Surf. Process. Landf. **33** (2008) 773. https://doi.org/10.1002/esp.1575
6 J. McKean, D. Tonina, C. Bohn, and C. W. Wright: J. Geophys. Res. Earth Surf. **119** (2014) 644. https://doi.org/10.1002/2013JF002897
7 H. F. Stockdon, A. H. Sallenger, J. H. List, and R. A. Holman: J. Coast. Res. **18** (2002) 502.
8 S. A. White and Y. Wang: Remote Sens. Environ. **85** (2003) 39. https://doi.org/10.1016/S0034-4257(02)00185-2
9 G. C. Guenther, A. G. Cunningham, P. E. LaRocque, and D. J. Reid: Proc. 20th EARSeL Symp. Workshop on Lidar Remote Sensing of Land and Sea (Dresden, Germany, 2000).
10 S. Pe'eri, C. Parrish, C. Azuike, L. Alexander, and A. Armstrong: Mar. Geod. **37** (2014) 293. https://doi.org/10.1080/01490419.2014.902880
11 J. Zhao, X. Zhao, H. Zhang, and F. Zhou: Remote Sens. **9** (2017) 710. https://doi.org/10.3390/rs9070710
12 T. Kogut and K. Bakuła: Remote Sens. **11** (2019) 1255. https://doi.org/10.3390/rs11111255

13 C. J. Legleiter, P. J. Kinzel, and B. T. Overstreet: Water Resour. Res. **47** (2011) W09531. https://doi.org/10.1029/2011WR010591
14 C. Mallet and F. Bretar: ISPRS J. Photogramm. Remote Sens. **64** (2009) 1. https://doi.org/10.1016/j.isprsjprs.2008.09.007
15 W. Wagner, A. Ullrich, V. Ducic, T. Melzer, and N. Studnicka: ISPRS J. Photogramm. Remote Sens. **60** (2006) 100. https://doi.org/10.1016/j.isprsjprs.2006.01.003
16 G. Mandlburger, M. Pfennigbauer, and N. Pfeifer: ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **II-5/W2** (2013) 175. https://doi.org/10.5194/isprsannals-II-5-W2-175-2013
17 R. Schwarz, G. Mandlburger, M. Pfennigbauer, and N. Pfeifer: ISPRS J. Photogramm. Remote Sens. **150** (2019) 1. https://doi.org/10.1016/j.isprsjprs.2019.02.002
18 C. Wang, Q. Li, Y. Liu, G. Wu, P. Liu, and X. Ding: ISPRS J. Photogramm. Remote Sens. **101** (2015) 22. https://doi.org/10.1016/j.isprsjprs.2014.11.005
19 S. Xing, D. Wang, Q. Xu, Y. Lin, P. Li, L. Jiao, X. Zhang, and C. Liu: Sens. **19** (2019) 5065. https://doi.org/10.3390/s19235065
20 T. Allouis, J. S. Bailly, Y. Pastol, and C. Le Roux: Earth Surf. Process. Landf. **35** (2010) 640. https://doi.org/10.1002/esp.1959
21 Z. Pan, C. Glennie, P. Hartzell, J. Fernandez-Diaz, C. Legleiter, and B. Overstreet: Remote Sens. **7** (2015) 5133. https://doi.org/10.3390/rs70505133
22 K. Richter, H.-G. Maas, P. Westfeld, and R. Weiß: PFG J. Photogramm. Remote Sens. Geoinf. Sci. **85** (2017) 31. https://doi.org/10.1007/s41064-016-0001-0
23 L. Wu, Y. Chen, Y. Le, Y. Qian, D. Zhang, and L. Wang: Int. J. Appl. Earth Obs. Geoinf. **128** (2024) 103770. https://doi.org/10.1016/j.jag.2024.103770
24 D. Mader, K. Richter, P. Westfeld, and H. G. Maas: ISPRS J. Photogramm. Remote Sens. **204** (2023) 145. https://doi.org/10.1016/j.isprsjprs.2023.08.014
25 K. Guo, W. Xu, Y. Liu, X. He, and Z. Tian: Remote Sens. **10** (2018) 35. https://doi.org/10.3390/rs10010035
26 H. Kim, M. Jung, J. Lee, and G. Wie: Appl. Sci. **13** (2023) 10939. https://doi.org/10.3390/app131910939
27 H. Kim and J. Lee: Remote Sens. **17** (2025) 3883. https://doi.org/10.3390/rs17233883
28 G. Liang, X. Zhao, J. Zhao, and F. Zhou: Remote Sens. **13** (2021) 3628. https://doi.org/10.3390/rs13183628
29 H. Kim, J. Lee, J. Kim, and H. Hur: J. Coast. Res. **116** (2024) 225. https://doi.org/10.2112/JCR-SI116-046.1
30 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 E. Duchesnay: J. Mach. Learn. Res. **12** (2011) 2825. https://dl.acm.org/doi/10.5555/1953048.2078195
31 L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone: Classification and Regression Trees (Wadsworth, Belmont, CA, USA, 1984).
32 L. Breiman: Mach. Learn. **45** (2001) 5. https://doi.org/10.1023/A:1010933404324
33 T. Chen and C. Guestrin: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2016) 785–794. https://doi.org/10.1145/2939672.2939785
34 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu: Proc. 31st Int. Conf. Neural Inf. Process. Syst. (2017) 3146–3154. https://dl.acm.org/doi/10.5555/3294996.3295074
35 I. Guyon, J. Weston, S. Barnhill, and V. Vapnik: Mach. Learn. **46** (2002) 389. https://doi.org/10.1023/A:1012487302797
36 H. Kim, G. H. Tuell, J. Y. Park, E. Brown, and G. Wie: J. Coast. Res. **91** (2019) 376. https://doi.org/10.2112/SI91-076.1
37 Korean Statistical Information Service (KOSIS): Water Quality in Coastal Waters. https://kosis.kr/statHtml/statHtml.do?orgId=146&tblId=DT_MLTM_1742 (accessed 2 November 2025).

About the Authors

Hyejin Kim received her B.S. and Ph.D. degrees from Seoul National University, South Korea, in 2001 and 2020, respectively. From 2021 to 2023, she was a chief researcher at Mokpo National University, South Korea. Since 2023, she has been a senior researcher at Konkuk University, South Korea. Her research interests are in ABL data processing, coastal surveying, and AI-based object detection and change analysis. (evervicky@konkuk.ac.kr)

Jaebin Lee received his B.S. degree from Yonsei University, South Korea, in 2000 and his Ph.D. degree from Seoul National University, South Korea, in 2008. Since 2009, he has been a professor at Mokpo National University. His research spans airborne bathymetric LiDAR, coastal and riverine mapping, AI-ready geospatial data processing, and digital-twin technologies. His interests include UAV/airborne LiDAR systems, coastal DEM generation, AI-based waveform analysis, and multimodal data fusion. (lee2009@mnu.ac.kr)