Ensemble Learning for Aerial LiDAR Point Cloud Semantic Segmentation: Exploring Model Complementarity and Combination Strategies

Alex Salvatierra[1 [0009−0005−8732−5581]], José Antonio Sanz[1 [0000−0002−1427−9909]], and Mikel Galar[1 [0000−0003−2865−6549]]

Department of Statistics, Computer Science and Mathematics and Institute of Smart Cities (ISC), Public University of Navarre (UPNA), Campus de Arrosadía s/n, Pamplona, 31006, Navarre, Spain
{alex.salvatierra, joseantonio.sanz, mikel.galar}@unavarra.es

Abstract. Deep learning models have demonstrated strong performance in semantic segmentation of aerial LiDAR point clouds, yet recent benchmarks reveal that state-of-the-art architectures achieving similar aggregate accuracy exhibit markedly different per-class behavior and error patterns. This diversity makes ensemble strategies a natural candidate for improving over individual models, although they remain unexplored in this domain. In this work, we present a systematic analysis of model complementarity and ensemble learning for aerial LiDAR semantic segmentation, evaluating four diverse state-of-the-art architectures across three large-scale datasets with distinct class taxonomies and geographic contexts. We quantify complementarity through oracle analysis and pairwise error overlap, and we propose an error exclusivity profile that characterizes how errors are distributed across models. We then evaluate a range of ensembles, from fixed combination rules to stacking. Our results confirm substantial complementarity across all benchmarks, with oracle gaps of 6–16 mIoU points, although current ensemble strategies recover only a fraction of this potential. A finer-grained evaluation shows that ensembles provide consistent gains across datasets and in several challenging classes, while also reducing the geometric severity of errors.

Keywords: Earth Observation · aerial LiDAR · point clouds · semantic segmentation · ensemble learning · model complementarity

1 Introduction

Semantic segmentation of aerial LiDAR point clouds has become a fundamental task in remote sensing, enabling the automated classification of large-scale 3D scenes into categories such as ground, vegetation, and buildings. The resulting labeled point clouds support the generation of essential geospatial products, including Digital Terrain Models and Digital Surface Models [22]. As the volume and spatial coverage of airborne laser scanning acquisitions continue to grow, automated methods have become essential for transforming raw point clouds into reliable geospatial products.

Deep learning has substantially advanced 3D point cloud segmentation since PointNet [1]. Recent architectures based on diverse design principles, including RandLA-Net [11], KPConv [29], Superpoint Transformer [25], and Point Transformer V3 [34], have achieved strong results on aerial LiDAR segmentation. Because these models encode geometry through different inductive biases, they often exhibit distinct class-level strengths and weaknesses. Recent spatially-aware evaluation further suggests that models with similar aggregate performance may fail on different subsets of points, indicating potentially complementary error patterns [28].

This complementarity naturally motivates the combination of multiple models to achieve better segmentation quality than any individual architecture alone. Ensemble learning is a well-established strategy for improving predictive performance by exploiting diversity across models [23, 26] and has been widely applied in 2D image segmentation [13] and other vision tasks, often yielding consistent improvements by leveraging complementary model behaviors. Aerial LiDAR point clouds, however, present unique characteristics that distinguish this domain from those where ensemble strategies have been previously studied: non-uniform point densities, massive spatial scales and severe class imbalance. These factors can amplify architecture-dependent error patterns, making model complementarity particularly relevant in this setting.

In this work, we present a systematic study of model complementarity and ensemble strategies for aerial LiDAR point cloud semantic segmentation. We first quantify complementarity through oracle analysis, pairwise error overlap, and a proposed error exclusivity profile, and then evaluate whether six ensemble strategies—from fixed combination rules [15] to stacking [32]—can exploit it, across four architecturally diverse models and three large-scale datasets: DALES [30], ECLAIR [21], and a proprietary dataset from the Pamplona metropolitan area. Evaluation extends beyond conventional mIoU through the spatially-aware framework defined in [28].

The main contributions of this work are:

1. A quantitative analysis of model complementarity for aerial LiDAR segmentation, revealing oracle gaps of 6–16 mIoU points and error exclusivity rates of 34–44% across three diverse benchmarks.
2. A systematic evaluation of six ensemble strategies, from fixed combination rules to stacking, across four architecturally diverse models and three datasets.

2 Related Work

2.1 Semantic Segmentation of Aerial Point Clouds

Semantic segmentation of 3D point clouds has advanced rapidly since PointNet [1] introduced direct processing of unordered point sets. Subsequent architectures addressed its limitations through hierarchical sampling [24], while later work explored several paradigms: MLP-based [3, 11], convolution-based [29, 33]

and graph-based [17, 31] approaches. More recently, transformer-based architectures have achieved state-of-the-art performance by applying attention mechanisms to point sets [25, 36].

While these models have been extensively evaluated on indoor and terrestrial benchmarks, the aerial LiDAR domain has received less attention. Aerial point clouds differ from terrestrial data in that they cover much larger spatial extents, exhibit stronger class imbalance, and are acquired from a nadir perspective producing distinct sampling and occlusion patterns. Benchmarks such as DALES [30], ECLAIR [21], FRACTAL [8], and YUTO [35] have expanded the availability of large-scale aerial data. Recent comparative studies [27] confirm that architecturally diverse models achieve competitive performance but exhibit distinct per-class strengths and weaknesses—yet whether these complementary error patterns can be exploited through ensemble strategies remains unexplored.

2.2 Ensemble Methods in Segmentation and Point Cloud Processing

Ensemble learning [14] improves predictive performance by combining multiple models through fixed combination rules [15] such as majority voting and probability averaging, or learned methods such as stacking [32], where a meta-learner is trained on base model outputs. Its effectiveness depends critically on base learner diversity: correlated errors yield limited benefit, while complementary errors create potential for substantial improvement [6, 16].

In aerial image segmentation, combining predictions from multiple networks has achieved strong results on 2D imagery benchmarks [19], establishing a precedent for model-level combination in remote sensing—albeit on 2D images rather than 3D point clouds. A related body of work addresses sensor-level fusion, combining complementary modalities such as hyperspectral imagery and LiDAR [5, 9]; however, this addresses complementarity at the input rather than the model level. Ensemble learning for 3D point cloud semantic segmentation remains largely unexplored, with only limited work in adjacent settings [2, 4]. To the best of our knowledge, it has not been studied for multi-class aerial point cloud semantic segmentation.

2.3 Limitations of Standard Segmentation Metrics

Standard evaluation of point cloud segmentation [10, 18] relies on overall accuracy (OA) and mean Intersection over Union (mIoU), which provide compact summaries of model performance but may obscure important differences in model behavior. In related domains, frequency-weighted IoU (fwIoU) has been adopted as a complementary metric to account for class imbalance [7].

Recent work [28] proposed a spatially-aware evaluation framework introducing distance-based metrics that quantify the geometric severity of misclassifications, and restricting evaluation to a *hard-points* subset—those misclassified by at least one compared model. Models with nearly identical full-test mIoU were shown to diverge by up to 29% in relative performance on these challenging regions, revealing that different architectures fail on different subsets of points.

This suggests that their error patterns may be complementary and potentially exploitable through ensembling. We adopt this framework alongside conventional mIoU, OA, and fwIoU, and investigate whether ensemble strategies can leverage these differences to improve segmentation quality.

3 Methodology

3.1 Complementarity Analysis

To quantify the extent to which different architectures make complementary errors, we combine an oracle analysis that estimates the upper bound of ensemble performance [15] with pairwise error overlap [16], measured here through the Jaccard index [12]. Furthermore, we propose an error exclusivity analysis to characterize how errors are distributed across models and to estimate what fraction of the oracle gap simple ensemble rules can realistically recover.

Let $X = \{p_i\}_{i=1}^{N}$ be the set of test points, $\mathcal{C}$ the set of semantic classes, and $\mathcal{M}$ the set of base models, with $M = |\mathcal{M}|$. Each point $p_i \in X$ is associated with a ground-truth label $y_i \in \mathcal{C}$; each model $m \in \mathcal{M}$ produces a predicted label $\hat{y}_i^{(m)} \in \mathcal{C}$ and a class-probability vector $\mathbf{P}_i^{(m)} \in [0,1]^{|\mathcal{C}|}$. The error set of model m is defined as

$$E^{(m)} = \left\{ p_i \in X \mid \hat{y}_i^{(m)} \neq y_i \right\}. \tag{1}$$

We first estimate the upper bound of achievable combination performance through an *oracle* combiner [15]. The oracle assigns the correct label to a point p_i whenever at least one base model predicts it correctly. The set of points correctly classified by the oracle is defined as

$$\mathcal{O} = \left\{ p_i \in X \mid \exists\, m \in \mathcal{M} : \hat{y}_i^{(m)} = y_i \right\}. \tag{2}$$

Oracle performance is obtained by computing the evaluation metrics over the full test set, where points in $\mathcal{O}$ are correct, and the *oracle gap* is defined as the difference between oracle performance and the best individual model.

To characterize similarity between any pair of models m_1 and m_2, we quantify pairwise error overlap [16] by computing the Jaccard index [12] between their error sets, which measures overlap as the ratio of shared errors to total errors across both models:

$$J(m_1, m_2) = \frac{|E^{(m_1)} \cap E^{(m_2)}|}{|E^{(m_1)} \cup E^{(m_2)}|}, \tag{3}$$

where lower values indicate more complementary error patterns, while higher values indicate that the two models tend to fail on the same points.

Beyond pairwise overlap, we propose an *error exclusivity profile* to characterize how errors are distributed across models and to estimate what fraction of the oracle gap simple ensemble rules can realistically recover. To this end, for every point p_i, we define its error count as

$$k_i = \sum_{m \in \mathcal{M}} \mathbb{I}\left[\hat{y}_i^{(m)} \neq y_i \right], \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The distribution of k_i over points with $k_i > 0$ defines the proposed error exclusivity profile. Low values of k_i indicate isolated errors and therefore higher local complementarity, whereas $k_i = M$ identifies points misclassified by all models, which cannot be recovered by recombining their predictions. Intermediate values correspond to partially shared errors, whose recoverability depends on the predicted classes and the combination rule. The class-conditional profile for class c is obtained by computing the same distribution over $\{p_i \in X \mid y_i = c,\ k_i > 0\}$.

3.2 Fixed Combination Rules

We evaluate three fixed combination strategies [15] that operate directly on the predictions of the trained base models and therefore require no additional training. Once the base predictions are available, these rules add negligible computational overhead. These methods differ in how they aggregate the outputs of the individual models.

Majority voting (F-MV) Each point is assigned the class receiving the most votes across the M base models:

$$\hat{y}_i^{\mathrm{MV}} = \arg\max_{c \in \mathcal{C}} \sum_{m \in \mathcal{M}} \mathbb{I}\Big[\hat{y}_i^{(m)} = c\Big]. \tag{5}$$

Ties are broken by summed softmax probability.

Probability averaging (F-PA) For each point p_i, we compute the arithmetic mean of the model probabilities:

$$\bar{\mathbf{P}}_i = \frac{1}{M} \sum_{m \in \mathcal{M}} \mathbf{P}_i^{(m)}, \tag{6}$$

where $\mathbf{P}_i^{(m)} \in [0,1]^{|\mathcal{C}|}$ is the class-probability vector produced by model m for point p_i. The final label is assigned as

$$\hat{y}_i^{\mathrm{PA}} = \arg\max_{c \in \mathcal{C}} \bar{P}_{i,c}, \tag{7}$$

where $\bar{P}_{i,c}$ denotes the averaged probability for class c.

Class-weighted averaging (F-CW) This strategy weights each model according to its per-class validation reliability. Let $w_c^{(m)}$ denote the validation IoU of model m for class c. The aggregated probability for class c at point p_i is

$$\bar{P}_{i,c}^{\mathrm{CW}} = \sum_{m \in \mathcal{M}} w_c^{(m)}\, P_{i,c}^{(m)}, \tag{8}$$

where $P_{i,c}^{(m)}$ is the probability assigned by model m to class c for point p_i. The final prediction is

$$\hat{y}_i^{\mathrm{CW}} = \arg\max_{c \in \mathcal{C}} \bar{P}_{i,c}^{\mathrm{CW}}. \tag{9}$$

3.3 Learned Combination via Stacking

While fixed combination rules use predefined aggregation schemes, stacking [32] learns a combiner from the outputs of the base models. For each point p_i, the meta-learner receives as input the concatenation of the class-probability vectors produced by the M base models:

$$\mathbf{x}_i = \left[\mathbf{P}_i^{(1)} \,\|\, \mathbf{P}_i^{(2)} \,\|\, \cdots \,\|\, \mathbf{P}_i^{(M)}\right], \tag{10}$$

where $\|$ denotes vector concatenation. This way, the final prediction can depend on the pattern of probabilities produced by the individual models rather than on a predefined voting or averaging scheme.

We evaluate three meta-learners with different modeling capacities: logistic regression (S-LR) as a linear combiner, and random forest (S-RF) and multilayer perceptron (S-MLP) as nonlinear alternatives. For the S-MLP, input features are standardized using z-score normalization before training. To prevent leakage, meta-learners are trained only on the validation split, with hyperparameters selected by 3-fold cross-validation and the final model retrained on the full validation split before test evaluation.

From a computational perspective, stacking introduces a lightweight learned combiner on top of the base predictions. The meta-learners operate only on $M|\mathcal{C}|$ softmax features per point, corresponding to 32, 44, and 20 input features for DALES, ECLAIR, and Tracasa-PNA20, respectively, rather than on raw point clouds, neighborhoods, or superpoints. In our experiments, the largest parametric meta-learner, S-MLP, contains only 13,000 trainable parameters, while S-LR contains at most 264; in comparison, the base backbones range from 0.21M parameters for SPT to 46.18M for PTv3. Therefore, the computational cost of stacking is dominated by training and running the four base models and storing their softmax outputs, rather than by the meta-learner itself.

3.4 Evaluation Protocol

We evaluate all individual models and ensembles on the test split using overall accuracy (OA), mean Intersection over Union (mIoU), and frequency-weighted IoU (fwIoU), which weights each class by its relative point frequency.

To assess performance beyond standard full-test evaluation, we adopt the evaluation framework of [28], which combines a subset of hard points with distance-based spatial error metrics. Specifically, the hard-points subset excludes points correctly classified by all models and focuses evaluation on challenging regions of the scenes. It is defined as

$$\mathcal{H} = \left\{ p_i \in X \,\middle|\, \exists\, m \in \mathcal{M},\ \hat{y}_i^{(m)} \neq y_i \right\}. \tag{11}$$

We report OA, mIoU, fwIoU, and per-class IoU on the full test set. On $\mathcal{H}$, which focuses evaluation on challenging regions and avoids dilution by trivially correct points, we additionally report OA, mIoU, and the macro Mean Distance Error

(mMDE). Specifically, MDE_c measures, for each misclassified point, the distance to the nearest correctly labeled point of the predicted class c: errors deep within a wrongly predicted region yield large distances, while boundary errors yield small ones. Its macro-average across classes, mMDE, summarizes the overall geometric severity of errors; lower values indicate spatially less severe misclassifications.

4 Experimental Setup

We evaluate on three large-scale aerial LiDAR datasets: DALES [30] (10 km^2, 505M points, 8 classes), ECLAIR [21] (10 km^2, 582M points, 11 classes), and Tracasa-PNA20, a proprietary dataset from the Pamplona metropolitan area (3.2 km^2, 194M points, 5 classes). All datasets have point densities in the 46–50 pts/m^2 range and cover diverse geographic contexts and class taxonomies.

We evaluate four base architectures selected for architectural diversity and strong performance: RandLA-Net [11], KPConv [29], Superpoint Transformer (SPT) [25], and Point Transformer V3 (PTv3) [34], spanning MLP-based, convo lution-based, superpoint-based, and transformer-based paradigms, respectively.

All models were trained independently for each dataset using publicly available implementations and hyperparameters from the original publications, adapted for the spatial scale and density of aerial data. Training was performed on NVIDIA RTX 6000 Ada GPUs with 48 GB of VRAM. Cross-entropy loss was used for all models, and the checkpoint with the lowest validation loss was selected for evaluation.

For all three datasets, the point clouds were partitioned into 50×50 m tiles for training and inference. In all cases, 3D coordinates were used as input features. Additional per-point attributes were dataset-dependent. For DALES, intensity was used in addition to the 3D coordinates. For ECLAIR, the input features included intensity, return number, number of returns, red, green, blue, and RGB average. For Tracasa-PNA20, the input features included intensity, return number, number of returns, red, green, blue, infrared, RGB average, and NDVI.

For each trained model and dataset, the full per-class softmax probability vector was stored for every validation and test point. These probability vectors serve as input to the fixed combination rules and to the stacking meta-learners.

5 Results & Discussion

5.1 Complementarity Analysis

To assess the degree of complementarity among the four base architectures, we first characterize pairwise redundancy through the Jaccard index between error sets. Table 1 summarizes the complementarity metrics across the three datasets.

Pairwise Jaccard analysis shows that the models fail on different sets of points. Mean pairwise Jaccard ranges from 0.46 on DALES to 0.53 on Tracasa-PNA20, indicating that roughly half of the errors made by any two models are not shared. The most complementary pairs are KPConv and SPT on both

Table 1. Summary of complementarity analysis across datasets.

Metric	DALES	ECLAIR	Tracasa-PNA20
Min pairwise Jaccard	0.39	0.39	0.49
Mean pairwise Jaccard	0.46	0.48	0.53
Best individual mIoU (%)	83.69 (PTv3)	72.65 (PTv3)	79.52 (PTv3)
Oracle mIoU (%)	90.35	88.42	85.67
Oracle gap (pp)	+6.66	+15.77	+6.15
$\|\mathcal{H}\|$ (% of test)	3.6	3.0	5.2
$k = 1$ (%)	43.7	43.4	33.9
$k = 2$ (%)	17.0	17.2	18.0
$k = 3$ (%)	14.7	14.0	17.7
$k = 4$ (%)	24.6	25.4	30.4

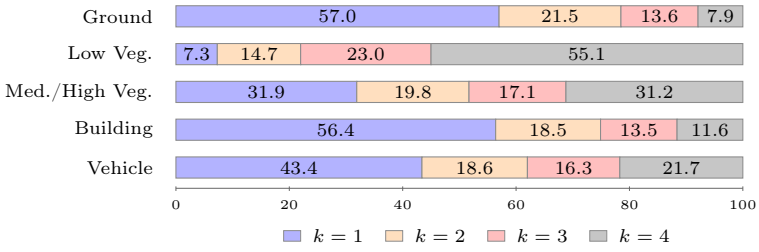

Fig. 1. Class-conditional error exclusivity profile on Tracasa-PNA20. Each bar shows the percentage of hard points misclassified by exactly k out of the four base models.

DALES and ECLAIR ($J = 0.39$), suggesting that the convolution-based and superpoint-based architectures fail on systematically different subsets of points.

To deepen this analysis, we examine the oracle gap and error exclusivity. The oracle gap, defined as the difference between oracle mIoU and the best individual-model mIoU, reaches +6.66 pp on DALES, +15.77 pp on ECLAIR, and +6.15 pp on Tracasa-PNA20, indicating substantial room for ensemble improvement. The larger gap on ECLAIR is consistent with its richer taxonomy (11 classes, versus 8 in DALES and 5 in Tracasa-PNA20), which creates more opportunities for architecture-specific failures to diverge.

To further assess how complementarity is distributed, we analyze the error exclusivity distribution, reported at the bottom of Table 1. This distribution partitions misclassified points by their error count k_i (Eq. 4). Points with $k = 1$ account for 43.7% of $\mathcal{H}$ in DALES, 43.4% in ECLAIR, and 33.9% in Tracasa-PNA20, corresponding to cases where only one model fails and the other three are correct, and thus recoverable by voting. However, these values are not expected gains over the best individual model: many such points may already be correctly classified by that model, contributing to the oracle gap rather than to the improvement over the strongest baseline. By contrast, 24.6–30.4% of errors form the irreducible core ($k = 4$), where all models fail simultaneously and no ensemble that only recombines their predictions can correct them.

A finer-grained per-class view is shown for Tracasa-PNA20 in Figure 1, revealing where ensemble potential is concentrated and where it is structurally

Table 2. OA, fwIoU, mIoU, and per-class IoU on the DALES test set.

				IoU							
Method	OA	fwIoU	mIoU	*Ground*	*Vegetation*	*Cars*	*Trucks*	*Power lines*	*Fences*	*Poles*	*Buildings*
KPConv	98.04	96.17	80.01	97.36	95.27	88.49	34.39	95.42	64.70	68.64	95.86
RandLA-Net	97.96	96.04	78.30	97.42	94.35	85.56	42.25	94.07	55.75	60.10	96.86
PTv3	98.23	96.55	83.69	97.64	94.86	89.28	49.87	**96.88**	68.55	**74.73**	97.70
SPT	97.90	95.90	83.03	97.01	94.18	89.71	**57.53**	94.81	64.14	69.81	97.06
F-MV	98.37	96.80	83.82	97.84	95.21	90.04	53.10	96.64	66.96	72.84	97.92
F-PA	98.37	96.81	83.99	97.84	95.22	90.12	53.62	96.68	67.40	73.07	**97.94**
F-CW	98.37	96.81	84.13	97.84	95.23	**90.21**	54.38	96.70	67.55	73.19	**97.94**
S-LR	**98.42**	**96.91**	**84.14**	97.96	**95.39**	89.92	53.40	96.59	68.48	73.54	97.86
S-MLP	98.37	96.83	82.61	**97.97**	95.28	85.67	46.05	96.63	**69.22**	72.30	97.77
S-RF	**98.42**	**96.91**	83.67	**97.97**	95.37	89.85	49.67	96.70	68.99	72.97	97.87
Oracle	99.12	98.26	90.35	98.86	97.38	94.16	74.78	98.19	79.60	80.93	98.90

Table 3. OA, fwIoU, mIoU, and per-class IoU on the ECLAIR test set.

				IoU									
Method	OA	fwIoU	mIoU	*Unassigned*	*Ground*	*Vegetation*	*Buildings*	*Noise*	*Trans. wires*	*Dist. wires*	*Poles*	*Trans. towers*	*Fence*	*Vehicle*
KPConv	98.65	97.35	72.62	15.05	97.33	97.45	**80.12**	65.51	**99.53**	93.30	65.24	74.88	66.15	44.31
RandLA-Net	98.32	96.70	58.29	1.83	96.74	96.86	60.85	34.38	97.96	73.02	28.15	50.72	48.08	52.62
PTv3	98.54	97.15	72.65	11.07	97.15	97.29	63.63	**69.61**	99.07	90.52	76.97	73.21	64.03	56.54
SPT	97.85	95.81	59.75	12.49	95.79	96.00	58.09	38.82	97.55	79.42	56.94	68.65	53.49	0.00
F-MV	98.64	97.33	74.14	14.28	97.31	97.45	71.69	65.82	99.51	**93.59**	72.88	78.51	68.96	55.56
F-PA	98.64	97.33	72.70	12.96	97.30	97.45	73.87	57.95	99.38	93.35	73.60	76.13	63.41	54.30
F-CW	98.64	97.33	**74.57**	12.39	97.30	97.45	72.20	68.23	99.31	92.40	76.34	77.95	65.41	**61.27**
S-LR	98.67	97.39	73.35	**17.30**	97.36	**97.52**	73.14	68.50	99.32	91.54	74.03	82.09	62.94	43.11
S-MLP	98.66	97.37	72.34	13.53	97.33	97.51	71.72	66.44	99.09	92.45	**78.70**	**85.64**	68.97	24.33
S-RF	**98.68**	**97.40**	73.86	16.42	**97.37**	97.51	79.14	67.45	99.25	91.53	76.17	74.25	**71.25**	42.15
Oracle	99.24	98.49	88.42	50.46	98.46	98.54	93.21	75.83	99.82	97.86	89.88	91.84	84.97	91.71

limited. *low vegetation* concentrates 55.1% of its errors at $k = 4$ and only 7.3% at $k = 1$, making it intrinsically resistant to ensemble recovery. By contrast, *ground* and *building* are dominated by $k = 1$ errors (57.0% and 56.4%), suggesting that ensemble combination should be more effective on these classes. These class-level patterns anticipate the behavior of the ensemble methods discussed next.

5.2 Ensemble Results on the Full Test Set

Tables 2–4 report results for all models and ensemble strategies across all three datasets. Comparisons are made against the best individual model for the dataset and metric under consideration. At the aggregate level, the strongest ensemble improves both OA and fwIoU on all three datasets, showing that ensemble fusion yields a net increase in correctly classified points. Although some fixed combination ensembles remain slightly below the best individual baseline in fwIoU on ECLAIR and Tracasa-PNA20, the best-performing combination strategy on each dataset consistently outperforms all individual models in both metrics.

McNemar's test [20] confirms that all differences between the best ensemble and the best individual model are statistically significant ($p < 0.05$); however, given that the test sets contain hundreds of millions of points, statistical signifi-

Table 4. OA, fwIoU, mIoU, and per-class IoU on the Tracasa-PNA20 test set.

				IoU				
Method	OA	fwIoU	mIoU	*Ground*	*Low Veg.*	*Med./High Veg.*	*Building*	*Vehicle*
KPConv	96.89	94.16	78.80	95.93	22.88	94.88	95.57	84.73
RandLA-Net	96.85	93.94	75.99	95.80	12.11	94.88	96.11	81.06
PTv3	97.01	94.67	79.52	96.18	**25.21**	95.33	96.82	84.08
SPT	96.46	93.12	72.61	95.12	0.76	94.26	95.87	77.04
F-MV	97.14	94.48	77.26	96.15	12.30	95.41	97.22	85.22
F-PA	97.12	94.43	76.95	96.12	10.64	95.41	97.23	85.37
F-CW	97.20	94.63	78.09	96.25	16.12	**95.47**	97.23	85.40
S-LR	**97.32**	**94.98**	**79.59**	**96.55**	22.90	95.36	**97.50**	**85.66**
S-MLP	97.30	94.90	78.71	96.54	21.65	95.26	97.41	82.71
S-RF	97.30	94.89	79.06	96.48	20.61	95.37	**97.50**	85.34
Oracle	98.41	96.89	85.67	97.89	43.65	97.38	99.04	90.41

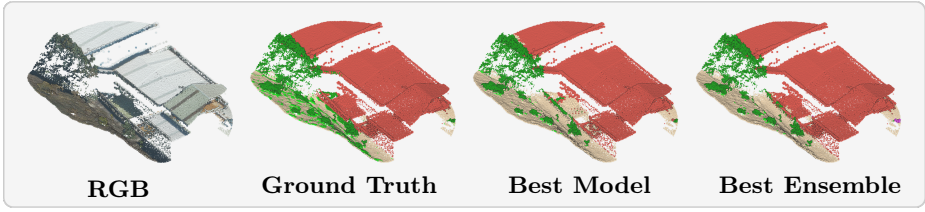

Fig. 2. Qualitative results on a Tracasa-PNA20 scene. From left to right: RGB input, ground truth, best individual model (PTv3), and best overall ensemble (S-LR). Each point is colored according to its semantic class: ground (beige), low veg. (bright green), med/high veg. (fern green), building (vermilion), vehicle (magenta).

cance is expected even for negligible differences and should not be interpreted as evidence of practical relevance. The odds ratio, defined as the ratio of points corrected by the ensemble to points where it introduces a new error, provides a more informative view, reaching up to 1.70 on Tracasa-PNA20, 1.59 on DALES, and 1.69 on ECLAIR, confirming a net positive effect across all benchmarks.

When considering full-test mIoU, which weights all classes equally, the analysis becomes much more sensitive to how corrections are distributed across classes. Under this metric, the best ensemble reaches 84.14 on DALES, 74.57 on ECLAIR, and 79.59 on Tracasa-PNA20, compared with 83.69, 72.65, and 79.52 for the best individual model. These gains remain small relative to the oracle gaps in Table 1. This more modest improvement is consistent with the fact that most points are already correctly classified by all individual models, while the underrepresented and geometrically complex classes that contribute equally to mIoU are often the hardest to improve.

The per-class results clarify this behavior. On DALES, the clearest gains appear in *cars* (90.21 vs. 89.71 best individual) and *buildings* (97.94 vs. 97.70), while *trucks* also improve from 49.87 to 54.38 despite a more mixed exclusivity profile. On ECLAIR, the largest improvements are observed in *vehicle* (61.27 vs. 56.54 best individual) and *trans. towers* (85.64 vs. 74.88), both characterized

Table 5. Performance on the hard-points subset $\mathcal{H}$ across datasets.

Method	DALES			ECLAIR			Tracasa-PNA20		
	OA	mIoU	mMDE	OA	mIoU	mMDE	OA	mIoU	mMDE
KPConv	45.11	29.14	1.35	55.08	51.48	1.04	40.46	30.22	0.92
RandLA-Net	43.03	23.71	1.69	43.90	22.97	1.87	39.71	27.24	0.98
PTv3	50.52	39.63	1.01	51.34	49.05	1.42	42.86	31.51	0.83
SPT	41.20	31.25	1.11	28.30	25.54	2.08	32.30	17.83	1.10
F-MV	54.25	40.43	**0.84**	54.70	**52.82**	0.87	45.25	33.74	0.65
F-PA	54.40	40.84	**0.84**	55.79	50.32	0.89	45.00	33.64	**0.64**
F-CW	54.46	41.11	**0.84**	55.59	52.62	**0.84**	46.43	34.68	0.65
S-LR	55.90	42.23	0.87	55.94	51.67	1.18	48.91	37.83	0.67
S-MLP	56.23	42.35	0.97	**56.77**	51.51	1.32	**48.97**	**38.15**	0.66
S-RF	**56.69**	**42.43**	**0.84**	56.69	52.14	0.94	48.76	37.91	0.66
Oracle	75.41	62.21	0.43	74.58	76.54	0.23	69.60	57.60	0.35

by high proportions of isolated errors and small irreducible cores. On Tracasa-PNA20, the behavior follows the exclusivity patterns discussed above: *building* and *vehicle*, dominated by $k = 1$ errors, improve under ensemble fusion, whereas *low vegetation*, with a highly unfavorable exclusivity profile, remains the main bottleneck and drops from 25.21 under PTv3 to as low as 10.64 under F-PA. Because *low vegetation* accounts for only 1.09% of points yet contributes equally to mIoU, this collapse offsets gains on other classes and helps explain why aggregate mIoU may remain close to, or even below, the best individual model despite a net increase in correct classifications.

Figure 2 illustrates these patterns on a representative Tracasa-PNA20 scene. The best ensemble visibly corrects several errors of the best individual model, particularly on *building* footprints and *ground* regions. By contrast, *low vegetation* errors—sparse and structurally ambiguous—largely persist after fusion, consistent with the high $k = 4$ exclusivity profile of this class.

Overall, full-test evaluation confirms that ensemble fusion is beneficial and consistent: odds ratios above 1.5 across all datasets confirm that ensembles correct more points than they disrupt, OA and fwIoU improve without exception, and per-class gains are observed in categories with favorable exclusivity profiles. How much this benefit reflects in mIoU depends strongly on class frequency and error structure. The most effective strategy varies by dataset: S-LR performs best on DALES and Tracasa-PNA20, while F-CW leads on ECLAIR. Yet even the best strategies recover only a modest fraction of the oracle gap, motivating the hard-points and distance-based analysis presented next.

5.3 Hard-Points and Distance-Based Evaluation

Table 5 reports mIoU, OA, and mMDE on the hard-points subset $\mathcal{H}$, focusing the analysis on the most challenging scene regions.

Under this evaluation, mIoU gains become substantially clearer than on the full test set. Among the learned ensembles, S-RF improves over PTv3 by +2.80 pp on DALES and +3.09 pp on ECLAIR, while S-MLP achieves the

largest gain on TRACASA-PNA20 with +6.64 pp. Fixed combination ensembles also improve over PTv3 on all three datasets, indicating that simple fusion rules already exploit part of the complementarity while learned meta-models capture additional structure from the base model probability patterns. These gains are substantially larger than under full-test mIoU, confirming that conventional evaluation absorbs much of the ensemble signal into trivially correct points.

The mMDE results further show that every ensemble method reduces the geometric severity of errors across all three datasets. In contrast to mIoU on $\mathcal{H}$, the strongest reductions are often achieved by fixed combination ensembles. On TRACASA-PNA20, mMDE decreases from 0.83 m to 0.64–0.65 m under fixed rules and 0.66–0.67 m under stacking. On DALES, PTv3 drops from 1.01 m to $\sim$0.84 m under the best fixed ensembles and S-RF. On ECLAIR, mMDE decreases from 1.42 m to 0.84–0.89 m under fixed methods and 0.94 m under S-RF. This pattern suggests that fixed rules tend to produce more conservative corrections that move residual errors closer to true class boundaries, even when they do not maximize the number of corrected points. This holds even for methods whose full-test mIoU gain is negligible or negative; for example, on TRACASA-PNA20, F-PA reduces mMDE from 0.83 m to 0.64 m despite decreasing full-test mIoU from 79.52 to 76.95.

Ensemble strategies therefore improve not only the quantity of correct predictions in challenging regions, but also the geometric quality of errors, which tend to occur closer to true class boundaries. This property is invisible to standard metrics yet directly relevant for geospatial applications where the spatial accuracy of segmentation determines the quality of derived geospatial products.

6 Conclusions

This paper presented the first systematic study of model complementarity and ensemble learning for aerial LiDAR point cloud semantic segmentation. We evaluated four architecturally diverse state-of-the-art models across three large-scale benchmarks spanning different geographic contexts, class taxonomies, and scene compositions; quantified complementarity through oracle analysis, pairwise error overlap, and the proposed per-class error exclusivity distribution; and assessed six ensemble strategies from fixed combination rules to learned meta-learners.

Three findings emerge consistently across datasets. First, different architectures exhibit substantial complementarity, as evidenced by oracle gaps of 6–16 mIoU points and error exclusivity rates of 34–44%, confirming that the four models fail on largely different point subsets; per-class exclusivity further reveals how each architecture encodes geometry differently. Second, ensemble strategies do exploit this complementarity: OA and fwIoU improve consistently, with per-class gains of up to +11 IoU points in categories with favorable exclusivity profiles. Under standard mIoU, however, only a fraction of this benefit is captured, since aggregate improvement depends strongly on how corrections distribute across classes, particularly under severe class imbalance. Third, evaluation on $\mathcal{H}$ shows that ensembles yield substantial mIoU improvements on challenging

regions, and that geometric error severity measured through mMDE decreases under every ensemble method and across all datasets, confirming that post-combination errors are spatially less severe and more contained, translating into higher-quality geospatial products.

These results show that ensemble fusion benefits in aerial LiDAR segmentation are consistent and reproducible, though the absolute gains remain modest: current strategies recover only a fraction of the available complementarity, and part of the improvement is additionally obscured by conventional metrics. More selective or context-aware fusion, particularly strategies capable of identifying when a minority prediction should override the majority, remains a promising direction for future work.

Acknowledgements Alex Salvatierra holds a predoctoral scholarship funded by the Tracasa Chair in Computer Science and Artificial Intelligence at the Public University of Navarre. This work was partially funded by project PID2022-136627NB-I00, funded by MCIN/AEI/10.13039/501100011033 and FEDER, EU; and by project PID2025-174319OB-I00, funded by MICIU/AEI/10.13039/5011 00011033 and FEDER, EU.

References

1. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–85 (2017). https://doi.org/10.1109/CVPR.2017.16
2. Chen, M., Feng, A., McCullough, K., Prasad, P.B., McAlinden, R., Soibelman, L.: 3D Photogrammetry Point Cloud Segmentation Using a Model Ensembling Framework. Journal of Computing in Civil Engineering **34**(6), 04020048 (2020). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000929
3. Choe, J., Park, C., Rameau, F., Park, J., Kweon, I.S.: PointMixer: MLP-Mixer for Point Cloud Understanding. In: Computer Vision – ECCV 2022. pp. 620–640. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19812-0_36
4. Ciou, T.S., Lin, C.H., Wang, C.K.: Airborne LiDAR Point Cloud Classification Using Ensemble Learning for DEM Generation. Sensors **24**(21) (2024). https://doi.org/10.3390/s24216858
5. Debes, C., Merentitis, A., Heremans, R., Hahn, J., Frangiadakis, N., van Kasteren, T., Liao, W., Bellens, R., Pižurica, A., Gautama, S., Philips, W., Prasad, S., Du, Q., Pacifici, F.: Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(6), 2405–2418 (2014). https://doi.org/10.1109/JSTARS.2014.2305441
6. Dietterich, T.G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning **40**(2), 139–157 (2000). https://doi.org/10.1023/A:1007607513941
7. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuScenes: A Large-Scale Benchmark for LiDAR Panoptic Segmentation and Tracking (2021). https://doi.org/10.48550/arXiv.2109.03805

8. Gaydon, C., Daab, M., Roche, F.: FRACTAL: An Ultra-Large-Scale Aerial Lidar Dataset for 3D Semantic Segmentation of Diverse Landscapes (2024). https://doi.org/10.48550/arXiv.2405.04634
9. Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A.: Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. IEEE Geoscience and Remote Sensing Magazine **7**(1), 6–39 (2019). https://doi.org/10.1109/MGRS.2018.2890023
10. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep Learning for 3D Point Clouds: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4338–4364 (2021). https://doi.org/10.1109/TPAMI.2020.3005434
11. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11105–11114 (2020). https://doi.org/10.1109/CVPR42600.2020.01112
12. Jaccard, P.: étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37**(142), 547 (1901). https://doi.org/10.5169/seals-266450
13. Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., Glocker, B.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Lect. Notes Comput. Sci. vol. 10670 LNCS, pp. 450–462. Springer Verlag (2018). https://doi.org/10.1007/978-3-319-75238-9_38
14. Krogh, A., Vedelsby, J.: Neural Network Ensembles, Cross Validation, and Active Learning. In: Advances in Neural Information Processing Systems. vol. 7. MIT Press (1994)
15. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Hoboken, NJ (2004). https://doi.org/10.1002/0471660264
16. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning **51**(2), 181–207 (2003). https://doi.org/10.1023/A:1022859003006
17. Landrieu, L., Simonovsky, M.: Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4558–4567 (2018). https://doi.org/10.1109/CVPR.2018.00479
18. Li, Y., Ma, L., Zhong, Z., Liu, F., Chapman, M.A., Cao, D., Li, J.: Deep Learning for LiDAR Point Clouds in Autonomous Driving: A Review. IEEE Transactions on Neural Networks and Learning Systems **32**(8), 3412–3432 (2021). https://doi.org/10.1109/TNNLS.2020.3015992
19. Marmanis, D., Wegner, J.D., Galliani, S., Schindler, K., Datcu, M., Stilla, U.: SEMANTIC SEGMENTATION OF AERIAL IMAGES WITH AN ENSEMBLE OF CNNS. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **III-3**, 473–480 (2016). https://doi.org/10.5194/isprs-annals-III-3-473-2016
20. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
21. Melekhov, I., Umashankar, A., Kim, H.J., Serkov, V., Argyle, D.: ECLAIR: A High-Fidelity Aerial LiDAR Dataset for Semantic Segmentation (2024). https://doi.org/10.48550/arXiv.2404.10699

22. Moore, I., Grayson, R., Ladson, A.: Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes **5**(1), 3–30 (1991). https://doi.org/10.1002/hyp.3360050103
23. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and Systems Magazine **6**(3), 21–44 (2006). https://doi.org/10.1109/MCAS.2006.1688199
24. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017). https://doi.org/10.48550/arXiv.1706.02413
25. Robert, D., Raguet, H., Landrieu, L.: Efficient 3D Semantic Segmentation with Superpoint Transformer. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17149–17158 (2023). https://doi.org/10.1109/ICCV51070.2023.01577
26. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review **33**(1-2), 1–39 (2010). https://doi.org/10.1007/s10462-009-9124-7
27. Salvatierra, A., Sanz, J.A., Gutiérrez, C., Galar, M.: Benchmarking Deep Learning Models for Aerial LiDAR Point Cloud Semantic Segmentation under Real Acquisition Conditions: A Case Study in Navarre (2026). https://doi.org/10.48550/arXiv.2603.22229
28. Salvatierra, A., Sanz, J.A., Gutiérrez, C., Galar, M.: Spatially-Aware Evaluation Framework for Aerial LiDAR Point Cloud Semantic Segmentation: Distance-Based Metrics on Challenging Regions (2026). https://doi.org/10.48550/arXiv.2603.22420
29. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.: KPConv: Flexible and Deformable Convolution for Point Clouds. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6410–6419 (2019). https://doi.org/10.1109/ICCV.2019.00651
30. Varney, N., Asari, V.K., Graehling, Q.: DALES: A Large-scale Aerial LiDAR Data Set for Semantic Segmentation (2020). https://doi.org/10.48550/arXiv.2004.11985
31. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic Graph CNN for Learning on Point Clouds. ACM Trans. Graph. **38**(5), 146:1–146:12 (2019). https://doi.org/10.1145/3326362
32. Wolpert, D.H.: Stacked generalization. Neural Networks **5**(2), 241–259 (1992). https://doi.org/10.1016/S0893-6080(05)80023-1
33. Wu, W., Qi, Z., Fuxin, L.: PointConv: Deep Convolutional Networks on 3D Point Clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9613–9622 (2019). https://doi.org/10.1109/CVPR.2019.00985
34. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point Transformer V3: Simpler, Faster, Stronger. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4840–4851 (2024). https://doi.org/10.1109/CVPR52733.2024.00463
35. Yoo, S., Ko, C., Sohn, G., Lee, H.: YUTO SEMANTIC: A LARGE SCALE AERIAL LIDAR DATASET FOR SEMANTIC SEGMENTATION. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLVIII-1-W2-2023**, 209–215 (2023). https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-209-2023
36. Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V.: Point Transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16239–16248 (2021). https://doi.org/10.1109/ICCV48922.2021.01595