Evaluating a Weighted Ensemble of Deep Learning Models for Individual Tree Crown Delineation from LiDAR Data

Dylan Gaspar[1], Baoxin Hu[1], Qian Li[1]

[1] Dept. of Earth and Space Science and Engineering, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada
(dylan07@my.yorku.ca, baoxin@yorku.ca, qian2018@yorku.ca)

Keywords: Individual Tree Crown Delineation, Canopy Height Model, Ensemble Method, Mask R-CNN, YOLO, U-Net

Abstract

This study investigates a weighted ensemble framework for individual tree crown (ITC) delineation using LiDAR-derived canopy height models (CHMs). Three deep learning models, Mask R-CNN, U-Net, and YOLO were first independently evaluated to establish the baseline performance under consistent training and evaluation conditions. A weighted ensemble was then constructed by combining model outputs through a voting-based fusion scheme, with an exhaustive search performed across multiple weight configurations to identify the ones that maximize common evaluation metrics. While certain weighting configurations yielded improvements in quantitative measures such as intersection over union (IoU), recall, F1 score, and accuracy relative to individual models, qualitative analysis revealed that these gains often coincided with substantial under-segmentation, manifested as large, merged crown regions. This discrepancy highlights the limitations of binary map voting for instance-level delineation and indicates that metric-driven ensemble optimization may not reliably reflect instance-level segmentation quality. The findings suggest that more expressive fusion strategies may be necessary for effective ensemble-based ITC delineation in future work.

1. Introduction

Accurate mapping and classification of forest structures is fundamental to ecological research and sustainable resource management. Individual tree crown (ITC) delineation plays a critical role in this process by enabling tree-level characterization, which supports precise analysis of forest composition and spatial configuration.

In the past couple decades, numerous methods have been developed for ITC delineation of canopy height models (CHMs) derived from light detection and ranging (LiDAR) data. Traditional approaches, such as watershed segmentation (Chen et al., 2006), local maxima filtering (Zhao and Popescu, 2007), and region-growing algorithms (Dalponte and Coomes, 2016), have been widely used, with varying success depending on canopy structure and species composition. More recently, convolutional neural network (CNN)–based deep learning models have been widely adopted to improve ITC delineation such as U-Net (Hu and Jung, 2021), YOLO (Straker et al., 2023), and Mask R-CNN (Li et al., 2025). These models represent different strategies for identifying and delineating ITCs from remote sensing data.

Despite their demonstrated success, these models exhibit distinct strengths and limitations when applied to complex forest canopies. U-Net produces accurate boundaries in open canopies, but struggles with overlapping crowns (Hu and Jung, 2021). YOLO offers efficient segmentation, but its precision can decrease in forest canopies characterized by irregular or atypical crown shapes (Straker et al., 2023). Mask R-CNN provides competitive accuracy relative to other approaches; however, it comes at the cost of higher computational complexity (Li et al., 2025). Collectively, these limitations suggest that no single model performs robustly across all forest configurations, motivating exploration of ensemble strategies that attempt to integrate the complementary strengths of multiple approaches.

To explore the concept further, in this study, a voting-based ensemble framework was evaluated for ITC delineation by combining Mask R-CNN, U-Net, and YOLO. The ensemble was assessed across a wide range of weight configurations to examine how different fusion strategies influenced common quantitative evaluation metrics. In addition to numerical performance, qualitative analysis was used to assess whether observed improvements reflect meaningful instance-level delineation or arise from artefacts such as crown merging and under-segmentation. Through this evaluation, the study aimed to identify both the potential and the limitations of binary voting-based ensemble strategies for ITC delineation.

2. Methods

2.1 Dataset

In this study, the three models were compared under consistent training conditions using LiDAR-derived CHMs collected from the study area located east of Sault Ste. Marie, Ontario, Canada (46◦33′43″ to 46◦34′03″N, 83◦25′10″ to 83◦25′25″W), as shown in Figures 1 and 2. This area is located in the Great Lakes-St. Lawrence Forest and has a mix of southern deciduous and northern boreal species. Common deciduous species include aspen (Populus tremuloides Michx.), white birch (Betula papyrifera Marsh.), and sugar maple (Acer saccharum), whereas most coniferous species are jack pine (Pinus banksiana Lamb.) and black spruce (Picea mariana Mill. BSP). Trees within the study area are densely distributed with multi-layered canopies of varying heights and ranging from 40 to 90 years in age.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-81-2026

Figure 1. Map of study site in Sault Ste. Marie, Ontario, Canada (46◦33′43″to 46◦34′03″N, 83◦25′10″to 83◦25′25″W).

LiDAR data for this study was acquired in August 2009 using a Riegl Q-560 scanner. The flight was approximately 200 m above ground, with a flight path that produced a high pulse density of approximately 40 pulses/m^2. . Rasterization was performed using a 0.15 m grid resolution, implemented with NumPy (1.26.0) and Rasterio (1.3.8). The Digital Surface Model (DSM) and Digital Elevation Model (DEM) were derived from the maximum elevation of vegetation and minimum elevation of ground points, respectively, within each 0.15 m × 0.15 m grid. The CHM was generated by computing the difference between the DSM and DEM in each grid, and post-processed to correct any invalid values, producing the grayscale seen in Figure 2 (Hyyppä and Inkinen, 1999; Hyyppä et al., 2001).

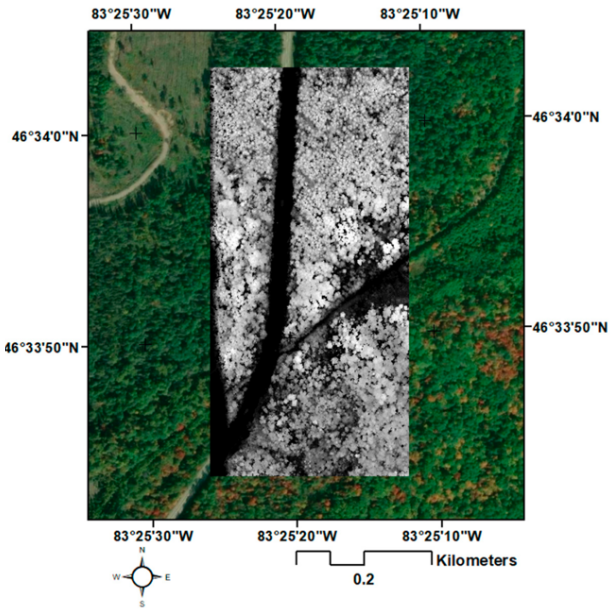

Figure 2. CHM of the study site (grayscale) overlapped with satellite image of the surrounding region.

A labelled reference dataset was created consisting of roughly 2500 manually delineated ITCs and was split into training and evaluation data. The training set covered approximately 75% of the area, with the evaluation set covering the remaining 25%. These percentages correspond to the approximate surface area covered by these splits, not the number of labelled reference crowns. For example, the number of segments used for evaluation is exactly 865, which corresponds to approximately 35% of total reference crowns. This reflects the higher crown density within the evaluation region compared with the training region, as shown in Figure 3.

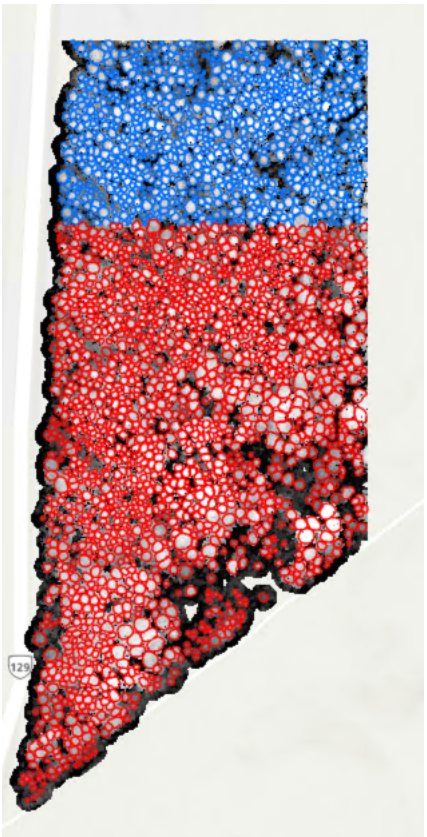

Figure 3. Canopy height model (CHM) of the study area showing manually delineated individual tree crowns and the spatial split of the dataset into training (red) and evaluation (blue) regions.

2.2 Methodology

In this study, a two-stage methodology was employed. Individual models were first implemented and trained under consistent conditions, after which an ensemble framework was constructed and evaluated across multiple weighting configurations. The following subsections describe individual model implementation, the ensemble framework, and the evaluation procedures.

2.2.1 Individual Model Implementation The Mask R-CNN model was built using a ResNet-50 backbone with a Feature Pyramid Network (FPN) (Lin et al., 2017) and initialized with COCO-pretrained weights (Lin et al., 2014). The classification and mask heads were modified to output a two-class prediction distinguishing tree crowns from background.

U-Net was implemented as a custom encoder-decoder semantic segmentation network with a ResNet-50 encoder initialized with ImageNet-pretrained weights (Deng et al., 2009). For U-Net, the predicted semantic probability map was post-processed using a watershed-based algorithm to derive individual crown instances.

YOLO was implemented as an instance segmentation model using the Ultralytics framework. Instead of a typical box detector, COCO-pretrained segmentation weights from YOLO26 (Ultralytics, 2026) were used to predict crown masks directly on

each tile. To ensure consistency with the other models, the predicted instance masks were converted into binary segmentation maps using Equation (1) for subsequent ensemble fusion.

$$M_i(x,y) = \begin{cases} 1 & if\ pixel\ part\ of\ polygon\ from\ model\ i \\ 0 & otherwise \end{cases} \quad (1)$$

where M_i = mask for model i
x, y = pixel coordinates

2.2.2 Ensemble Framework In the second stage, an ensemble framework was constructed in which individual models were treated as supporting agents and combined using a weighted voting strategy, with each model assigned a corresponding weight. A weighted support map is then computed using the equation seen below.

$$E(x,y) = \sum_{i=1}^{3} w_i \cdot M_i(x,y) \quad (2)$$

where E = ensemble support value at pixel (x,y)
w_i = weight for model i, where $\sum w_i = 1$

Finally, this continuous valued mask is converted into a binary one using a threshold filter. The threshold, T, was fixed at 0.5 through all iterations.

$$E_{final}(x,y) = \begin{cases} 1 & if\ E(x,y) > T \\ 0 & otherwise \end{cases} \quad (3)$$

where E_{final} = final binary segmentation mask
T = 0.5

To systematically explore the effect of model weighting, an exhaustive grid search was performed over all possible weight combinations at increments of 0.1. For each configuration, ensemble outputs were evaluated using selected quantitative metrics. The resulting performance values were recorded to enable comparison of ensemble behaviour across weighting configurations and evaluation criteria.

2.2.3 Evaluation Metrics Evaluations of all three base models and the ensemble-based model(s) are carried out using a collection of quantitative and qualitative measures. Quantitative metrics include intersection over union (IoU) and basic classification model measures such as precision, recall, F1 score, and accuracy, which collectively provides insight into spatial overlap, detection behaviour, and class prediction tendencies. Qualitative evaluation was performed through visual inspection of predicted crown delineations to assess instance separation, boundary accuracy, and common failure modes such as over-merging and under-segmentation.

In addition to segmentation metrics, computational efficiency was assessed by measuring inference time for each individual model. An accuracy-per-time (APT) score was computed to characterize the trade-off between segmentation accuracy and inference speed, providing a relative measure of model efficiency (Teerapittayanon et al., 2017). Inference time and APT was used solely for comparative analysis of individual models and was not applied in the ensemble evaluation.

3. Results

3.1 Individual Model Performances

		Values		
	Metric	Mask R-CNN	U-Net	YOLO
Counts	Predictions	582	440	833
	True Positives	477	200	685
	False Positives	105	240	148
	False Negatives	388	665	180
Rates	IoU	0.681	0.642	0.698
	Precision	0.820	0.455	0.822
	Recall	0.551	0.231	0.792
	F1 Score	0.659	0.307	0.807
	Accuracy	0.492	0.181	0.676
	Time (s)	12.1	2.90	4.68
	APT (acc/s)	0.041	0.063	0.144

Table 1. Performance of Individual Models.

As reported in Table 1, YOLO demonstrated stronger performance over Mask R-CNN and U-Net across most of the measures. YOLO achieved the highest IoU (0.698), precision (0.822), recall (0.792), F1 score (0.807), and accuracy (0.676). These values reflect YOLO’s higher number of true positives (685) and substantially lower number of false negatives (180) compared with the other models. In addition, YOLO exhibited faster inference time (4.68 s) than Mask R-CNN and achieved the highest accuracy-per-time (APT) score (0.144), indicating a favorable balance between segmentation performance and computational efficiency. Figure 4 illustrates YOLO’s predicted crown delineations in a densely clustered canopy section, where the model tends to preserve separate crown instances, contributing to the higher detection-based metrics reported in Table 1.

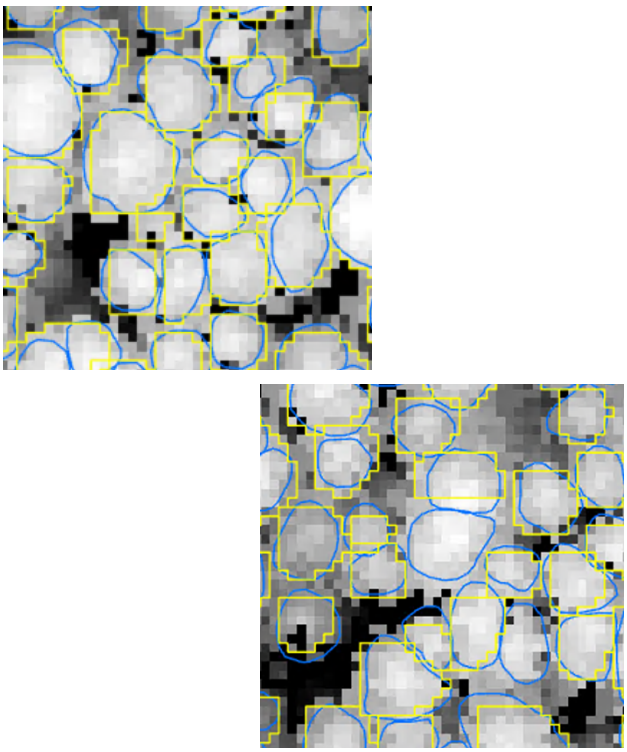

Figure 4. Illustration of individual tree crowns predicted by YOLO (yellow) overlaid with the manually labelled reference crowns (blue) in a densely clustered canopy section.

Mask R-CNN achieved relatively high precision (0.820) and IoU (0.681) among the evaluated models, as shown in Table 1. The model generated 582 predicted instances, with 477 true positives and 105 false positives, indicating that predicted crowns were generally well aligned with reference crowns when detections occurred. However, Mask R-CNN also exhibited a high number of false negatives (388), resulting in comparatively lower recall (0.551), accuracy (0.492), and F1 score (0.659). Figure 5 presents qualitative examples of Mask R-CNN outputs, illustrating regions where crown predictions are omitted in dense canopy areas despite relatively precise delineation of detected crowns.

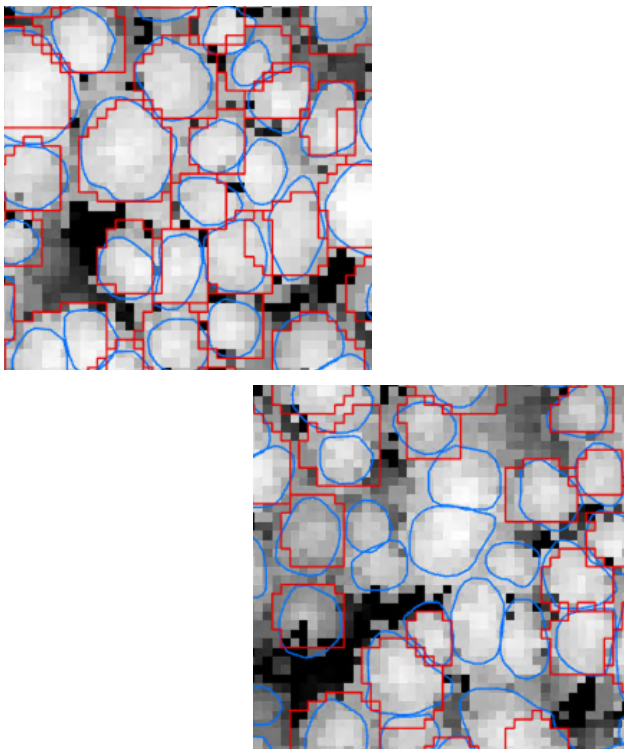

Figure 5. Qualitative examples of Mask R-CNN (red). Top: scenario where Mask R-CNN does not have the same false positive that YOLO produced. Bottom: example of considerable omission bias.

U-Net exhibited substantially different behavior from the instance-based models, producing fewer predicted crown instances (440) and lower detection-based performance across all evaluated metrics (Table 1). The model yielded 200 true positives alongside 240 false positives and 665 false negatives, resulting in low precision (0.455), recall (0.231), F1 score (0.307), and accuracy (0.181), despite a moderate IoU value (0.642). Qualitative inspection of U-Net outputs following watershed-based post-processing revealed extensive crown merging in densely forested regions. Figure 6 illustrates examples where large, connected regions span multiple individual crowns, highlighting the dominant segmentation behavior underlying the reported quantitative results.

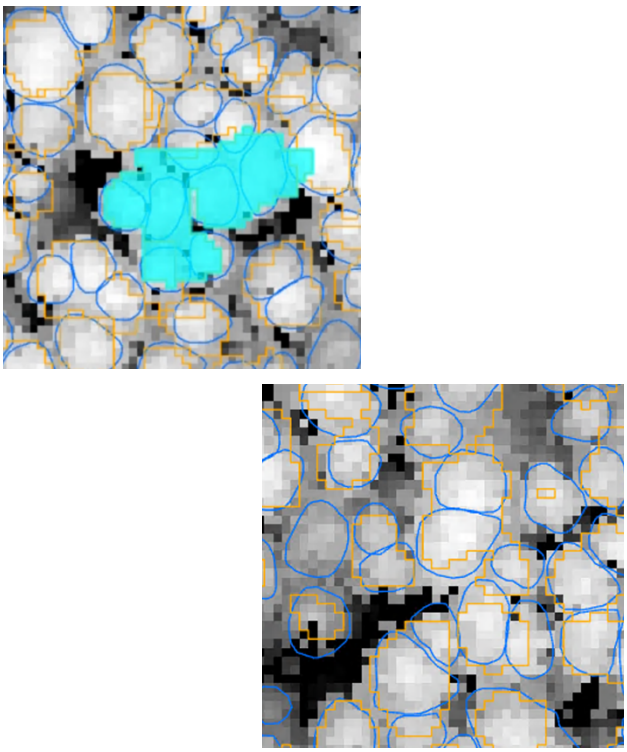

Figure 6. Qualitative examples of U-Net (orange). Top: highlighted area showing extreme case of under-segmentation. Bottom: more under-segmentation, but less extreme.

3.2 Ensemble Model Performance

The script that iterates through all possible weighting configurations was employed to identify the configurations that provide the highest values in IoU, precision, recall, F1 score, and accuracy. Table 2 shows the findings for each metric.

		Weights		
Metric	Highest	Mask R-CNN	U-Net	YOLO
IoU	0.916	0.5	0.0	0.5
Precision	0.843	0.5	0.0	0.5
Recall	0.881	0.5	0.0	0.5
F1 Score	0.862	0.5	0.0	0.5
Accuracy	0.757	0.5	0.0	0.5

Table 2. Weighting configurations yielding highest metrics.

Table 2 shows that all the top-performing ensemble configurations relied on equal combinations of YOLO and Mask R-CNN, with no optimal configuration needing contributions from U-Net.

Results in Table 3 below show that this optimal fusion of YOLO and Mask R-CNN produced the highest values of IoU (0.916), precision (0.843), recall (0.881), F1 score (0.862) and accuracy (0.757). This ensemble framework also produces the most predictions (897) and true positives (762), with the fewest false negatives (103). Figure 7 shows qualitative results for a small region with no false positives or false negatives.

	Metric	Values
Counts	Predictions	897
	True Positives	762
	False Positives	142
	False Negatives	103
Rates	IoU	0.916
	Precision	0.843
	Recall	0.881
	F1 Score	0.862
	Accuracy	0.757

Table 3. Performance of ensemble (green) with $w_{Mask\ R\text{-}CNN} = 0.5$ and $w_{YOLO} = 0.5$.

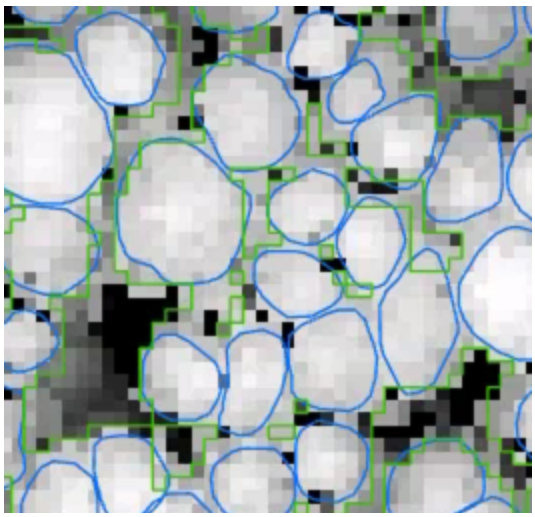

Figure 7. Qualitative results of ensemble with $w_{Mask\ R\text{-}CNN} = 0.5$ and $w_{YOLO} = 0.5$

A direct comparison between the best performing individual model (YOLO) and the optimal ensemble configuration is shown in Table 4, where an increase is seen in IoU, precision, recall, F1 score, and accuracy.

	Values	
Metric	YOLO	Ensemble
IoU	0.698	0.916
Precision	0.822	0.843
Recall	0.792	0.881
F1 Score	0.807	0.862
Accuracy	0.676	0.757

Table 4. Comparing best individual model (YOLO) performance to ensemble performance.

4. Discussion

4.1 Sensitivity to Weighting and Threshold Choices

A key limitation of this proposed ensemble framework and specific weighting strategy was its strong dependence on the selected weight values. Since this process combines binary model outputs, where each pixel has a value of exclusively 1 or 0, the influence of each model is governed mainly by whether its assigned weight can sufficiently surpass the vote threshold on its own. This helps explain why optimized weight combinations repeatedly selected weights of 0.5, consistent with the voting threshold being set to 0.5.

4.2 Mismatch Between Metric Optimization and Delineation Quality

In addition, the fusion strategy employed tended to merge adjacent crowns from differing models into much larger connected regions. This led to extreme under-segmentation, where multiple reference tree crowns were covered by a single predicted segment as seen in Figure 8. As a result, the ensemble may produce visually severe under-segmentation despite relatively strong quantitative performance.

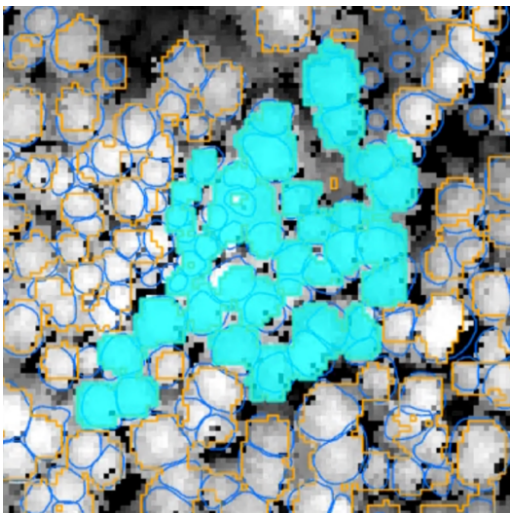

Figure 8. Extreme case of under-segmentation and over-merging ($w_{Mask\ R\text{-}CNN} = 0.33$, $w_{U\text{-}Net} = 0.33$, $w_{YOLO} = 0.33$).

The discrepancy between improved metrics and degraded instance-level delineation is caused by how these metrics reward spatial overlap rather than segmentation correctness. The metrics used in this study favour larger regions because they increase the proportion of overlap between predicted and labelled reference segments. When multiple adjacent crowns are merged into a single prediction, high overlap can still be achieved with several labelled reference crowns simultaneously. The many-to-one mapping behaviour of these metrics limits the penalty of under-segmentation by allowing a single over-merged prediction to maintain high overlap with multiple labelled reference segments. This effect is amplified in dense canopies, where closely spaced trees increase the likelihood of merging.

4.3 Limitations of Binary Voting

Beyond the metric-related effects mentioned above, the observed patterns reflect a broader limitation of binary voting-based ensemble methods. In some configurations, ensembles may reinforce agreement rather than resolve disagreements. Binary maps discard valuable information such as confidence and probability, preventing the fusion from distinguishing between overlapping but distinct predictions, which encourages combination into larger predicted regions. This behaviour suggests that binary voting ensembles are limited in their ability to improve segmentation tasks where instance separation is critical.

4.4 Implications for Ensemble Strategies in ITC Delineation

The findings of this study highlight important considerations for designing ensemble strategies in ITC delineation tasks. While weighted combinations of binary masks from multiple models can improve quantitative metrics, these do not necessarily

translate to greater instance-level segmentation. In particular, the fusion of binary masks may be unsuitable for ITC delineation, as their tendency to under-segment and over merge tree crowns limits their effectiveness for applications requiring precise delineation. Future voting-based ensemble strategies need to consider utilizing confidence or probability scores instead of binary values.

Despite these limitations, binary voting-based ensembles may still be appropriate in some contexts. For example, they may be effective for general tree presence mapping where precise delineation of individual crowns is not necessary. Also, sparse forest environments with distant trees reduce the likelihood of over merging, making binary fusion more reliable.

5. Conclusion

This study evaluated the effectiveness of a weighted ensemble framework for ITC delineation using LiDAR-derived CHMs by combining Mask R-CNN, U-Net, and YOLO through a voting-based fusion strategy. By systematically exploring ensemble weight configurations and comparing ensemble outputs with those of the individual models, the study examined how binary fusion strategies influence commonly used quantitative evaluation metrics.

The ensemble framework demonstrated that combining model outputs could lead to improvements in several quantitative metrics, including IoU, precision, recall, F1 score, and accuracy. The highest performing configuration consistently relied on a combination of YOLO and Mask R-CNN, achieving the strongest overall metric values. However, qualitative analysis revealed that these configurations frequently produced large, under-segmented regions that merged multiple crowns into single predicted segments. This highlights that improvements in quantitative metrics do not necessarily correspond to improved instance-level tree crown separation and reveal a key limitation of binary voting-based ensemble approaches for ITC delineation. While such ensembles can enhance commonly used quantitative metrics, they tend to favor spatial overlap and often exhibit a bias toward significant under-segmentation, resulting in merged crown predictions.

Among the individual models, YOLO demonstrated the strongest overall performance across detection-based metrics, while Mask R-CNN provided competitive spatial precision with higher omission rates. U-Net showed limited effectiveness in this context, with substantially lower performance across most evaluation measures. These differences highlight the variability in how individual models balance detection performance and crown separation. This variability also motivates the continued exploration of ensemble approaches that can better integrate complementary model behaviors while addressing instance-level delineation challenges.

In future work, alternative ensemble fusion strategies that utilize confidence scores or probability maps rather than binary outputs, will be investigated, as continuous representations may allow for more nuanced integration of model predictions and reduce sensitivity to weight selection. In addition, further validation across different forest conditions and datasets will be conducted to assess the generality of the observed ensemble behavior.

Acknowledgements

The authors are grateful for financial support provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and York University (Research-At-York program).

References

Chen, Q., Baldocchi, D., Gong, P., & Kelly, M. (2006). Isolating individual trees in a savanna woodland using small-footprint LiDAR data. *Photogrammetric Engineering & Remote Sensing, 72*(8), 923–932. https://doi.org/10.14358/PERS.72.8.923

Dalponte, M., & Coomes, D. A. (2016). *Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data*. *7*(10), 1236–1245. https://doi.org/10.1111/2041-210x.12575

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Hu, B., & Jung, W. (2021). Individual Tree Crown Delineation From High Spatial Resolution Imagery Using U-Net. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B3-2021*, 61–66. https://doi.org/10.5194/isprs-archives-xliii-b3-2021-61-2021

Hyyppä, J., & Inkinen, M. (1999). Detecting and estimating attributes for single trees using laser scanner. *The Photogrammetric Journal of Finland, 16*(2), 27–42. https://cir.nii.ac.jp/crid/1571698600725687936

Hyyppä, J., Kelle, O., Lehikoinen, M., & Inkinen, M. (2001). A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. IEEE Transactions on Geoscience and Remote Sensing, 39(5), 969–975. https://www.scopus.com/pages/publications/0035332402

Li, Q., Hu, B., Shang, J., & Remmel, T. K. (2025). Two-Stage Deep Learning Framework for Individual Tree Crown Detection and Delineation in Mixed-Wood Forests Using High-Resolution Light Detection and Ranging Data. *Remote Sensing*, *17*(9), 1578. https://doi.org/10.3390/rs17091578

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *ArXiv:1612.03144 [Cs]*. https://arxiv.org/abs/1612.03144

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, *8693*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Straker, A., Puliti, S., Breidenbach, J., Christoph Kleinn, Pearse, G., Astrup, R., & Magdon, P. (2023). Instance segmentation of individual tree crowns with YOLOv5: A comparison of approaches using the ForInstance benchmark LiDAR dataset. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *9*, 100045–100045. https://doi.org/10.1016/j.ophoto.2023.100045

Teerapittayanon, S., McDanel, B., & Kung, H. T. (2017). BranchyNet: Fast Inference via Early Exiting from Deep Neural

Networks. *ArXiv:1709.01686 [Cs]*. https://arxiv.org/abs/1709.01686

Ultralytics. (2026). Ultralytics YOLO (Version 26) https://github.com/ultralytics/ultralytics

Zhao, K., & Popescu, S. (2007). *Hierarchical Watershed Segmentation Of Canopy Height Model For Multi-Scale Forest Inventory*. Retrieved October 9, 2025, from https://www.isprs.org/proceedings/xxxvi/3-w52/final_papers/zhao_2007.pdf