Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 4, November, Pp. 535-546

Modified U-Net for leaf segmentation of Eucalyptus pellita seedlings in open nursery environments

Tegar Alami[1], Yeni Herdiyeni*[1], Wisnu Ananta Kusuma[2], Budi Tjahjono[3], Iskandar Zulkarnaen Siregar[4]
Artificial Intelligence Study Program, IPB University, Bogor, West Java, Indonesia[1]
Bioinformatics Study Program, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, West Java, Indonesia[2]
APP Academy, Asia Pulp and Paper, Pekanbaru, Riau, Indonesia[3]
Department of Silviculture, Faculty of Forestry and Environment, IPB University, Bogor, West Java, Indonesia[4]

Article Info

Keywords:
Leaf Segmentation, Modified U-Net, ResNet-50, Eucalyptus, Natural Environment, Deep Learning

Article history:
Received: June 05, 2025
Accepted: September 17, 2025
Published: November 01, 2025

Cite:
T. Alami, Y. Herdiyeni, W. A. Kusuma, B. Tjahjono, and I. Z. Siregar, "Modified U-Net for Leaf Segmentation of Eucalyptus pellita Seedlings in Open Nursery Environments", *KINETIK*, vol. 10, no. 4, Nov. 2025. https://doi.org/10.22219/kinetik.v10i4.2349

*Corresponding author.
Yeni Herdiyeni
E-mail address:
yeni.herdiyeni@apps.ipb.ac.id

Abstract

This study addressed leaf segmentation in open nursery environments for Eucalyptus pellita seedlings, where fluctuating illumination, cluttered backgrounds, and overlapping foliage had hindered reliable monitoring at operational scale. We proposed a Modified U-Net that integrated a ResNet-50 encoder for high-resolution feature extraction, L2 regularization in the decoder to improve generalization, and a composite binary cross-entropy plus Dice loss to balance pixel-level accuracy with shape conformity. We assembled 2,424 RGB images from an operational nursery and evaluated three architectures (Modified U-Net as the primary model, SegNet, and DeepLabv3+) under cloudy, sunny, and scorching illumination. We conducted inference at native resolution and summarized per-image metrics using medians with interquartile ranges, followed by nonparametric significance testing. The Modified U-Net consistently outperformed the baselines across all scenarios, achieving median Dice coefficients of 0.872 (cloudy), 0.841 (sunny), and 0.854 (scorching), with corresponding Intersection over Union values of 0.773, 0.725, and 0.745. A Kruskal-Wallis test on per-image Dice and Intersection over Union yielded no significant differences across lighting conditions (H = 4.012, p = 0.1345), indicating stable performance under natural illumination variability. Qualitative overlays revealed localized errors, including glare-induced false positives in sunny scenes and shadow-related artifacts under scorching light, which did not materially shift global overlap distributions. We concluded that the proposed architecture delivered robust, high-fidelity segmentation in realistic nursery conditions and provided a practical basis for field deployment, with further gains expected from glare- and shadow-aware augmentation and lightweight optimization for near real-time inference on edge devices.

1. Introduction

Indonesia is a megadiverse country with a substantial contribution to the forestry sector, particularly through the management of Industrial Forest Plantations (IFP) [1]. According to Statistics Indonesia [2], national roundwood production reached 68.22 million m³, dominated by *Acacia* (56.61%, approximately 38.6 million m³) and *Eucalyptus* (41.68%, approximately 28.44 million m³). The increasing use of *Eucalyptus pellita* is closely linked to a policy-driven shift from peatlands to mineral soils [3]. Among *Eucalyptus* species, *Eucalyptus pellita* is widely cultivated by the pulp and paper industry due to its rapid growth and broad ecological adaptability [4]. In IFP operations, the nursery phase is critical because early biotic and abiotic disturbances can significantly reduce productivity and quality in subsequent growth stages [5]. Accurate and continuous seedling health monitoring is therefore essential for effective and sustainable silvicultural practices [6].

Conventional monitoring methods that rely on manual visual inspection face multiple limitations, including inefficient use of time, dependence on skilled personnel, low reproducibility, and inconsistent accuracy at nursery scale [7]. These limitations have encouraged the development of AI-based approaches that leverage digital imagery acquired by unmanned aerial vehicles (UAVs), RGB cameras, or handheld optical devices [8]. In recent years, deep learning-based image segmentation has become a robust solution in computer vision for agriculture, enabling precise detection and mapping of plant structures [9]. However, studies using U-Net variants, SegNet, and DeepLab families, while effective in controlled settings, often experience performance drops under fluctuating illumination and cluttered backgrounds, especially at high resolution [10].

Leaf segmentation in open nursery environments remains particularly challenging [11]. Variations in lighting (cloudy, sunny, and scorching), complex backgrounds, and overlapping foliage frequently reduce model performance [12]. Photometric inconsistency caused by shadows, specular highlights, and overexposure blurs object boundaries and triggers segmentation errors [13]. These early-stage errors can propagate and reduce the accuracy of downstream tasks, including defoliation assessment and disease detection [14]. Despite the growing number of studies on leaf segmentation in agricultural research, systematic investigations focusing on *Eucalyptus pellita* seedlings in open nursery conditions are still limited [15]. This gap is critical because seedlings represent the most vulnerable stage in IFP operations, and the lack of robust segmentation models constrains the early detection and management of health-related risks in forestry nurseries [6].

To address these challenges, this study proposes a Modified U-Net specifically optimized for outdoor leaf segmentation. The model incorporates three targeted enhancements: a ResNet-50 encoder to strengthen high-resolution feature extraction, L2 regularization in the decoder to improve generalization, and a composite Binary Cross-Entropy and Dice loss to balance pixel-level accuracy with shape conformity under class imbalance. These modifications are systematically evaluated under three natural lighting scenarios (cloudy, sunny, and scorching) using a large dataset of *Eucalyptus pellita* seedlings. The novelty of this work lies in combining architectural improvements with rigorous field-based validation on a high-value forestry species. Unlike previous studies that emphasize laboratory or simplified settings, this research explicitly addresses the gap in robust segmentation for *Eucalyptus pellita* seedlings under complex real-world nursery environments, providing a framework with both scientific novelty and operational relevance.

This work makes three contributions. First, we introduce a Modified U-Net tailored for outdoor leaf segmentation on *Eucalyptus pellita* seedlings, combining a ResNet-50 encoder for high-resolution feature extraction, L2-regularized decoder layers to improve generalization, and a composite Binary Cross-Entropy plus Dice loss that balances pixel accuracy with shape fidelity under class imbalance. Second, we deliver a field-validated evaluation under three natural lighting scenarios (cloudy, sunny, and scorching) on a large, real-world nursery dataset, reporting quantitative and qualitative results with rigorous statistics (Kruskal–Wallis with epsilon-squared effect size) and median [IQR] summaries for transparent, reproducible comparison. Third, we demonstrate consistent gains over SegNet and DeepLab families across lighting conditions on Dice and IoU, with no statistically significant differences across illumination and negligible effect sizes, and we distill deployment-oriented guidance (e.g., illumination-aware augmentation and acquisition practices) to support robust nursery operations.

2. Research Method

The research methodology was divided into two main phases. The first phase was data preparation, which included image acquisition, preprocessing, annotation, and dataset splitting. This phase ensured that the dataset captured the variability of natural lighting conditions in the field. The second phase was model development and evaluation, which involved the design and training of three architectures (Modified U-Net as the primary model, with SegNet and DeepLabv3+ as baselines), model optimization, and quantitative as well as statistical performance evaluation. A schematic representation of the workflow is presented in Figure 1, showing the sequential process from raw image collection to model evaluation. This approach was designed to address existing research gaps and to meet the operational needs of large-scale nursery management.

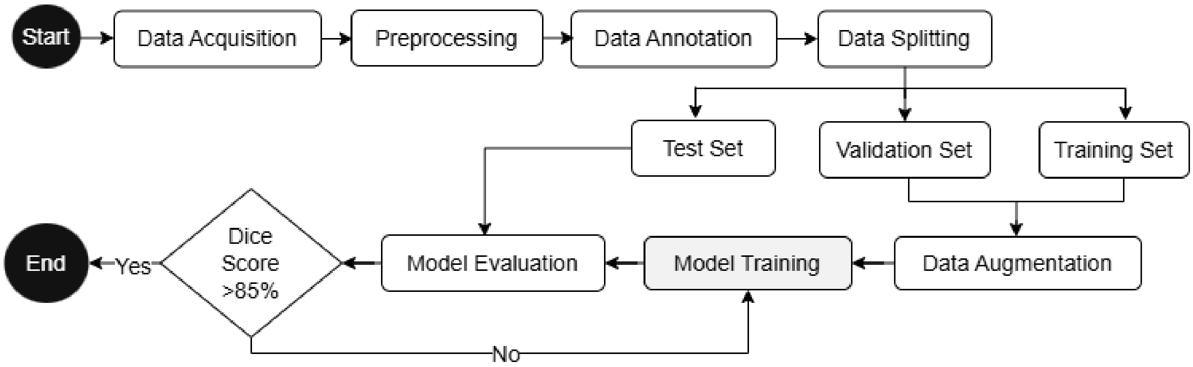

Figure 1. Workflow of the Proposed Leaf Segmentation Model using the Modified U-Net

2.1 Data Preparation

Image acquisition was conducted on *Eucalyptus pellita* seedlings in an intensively managed open nursery to capture operational field conditions representative of Industrial Forest Plantations [16]. Images were collected using a Logitech RGB USB 1080p camera connected to an Intel NUC i7 mini-PC. The camera was mounted in a fixed position with consistent distance and nadir angle using a boom sprayer rig, which minimized perspective distortion and reduced vibration or operator shadows. Acquisition was performed every morning between 07:00 and 08:00 before irrigation to ensure dry leaves, avoid water-induced reflections, and maintain relatively stable natural illumination. All images were recorded at their native resolution of 1920 × 1080 pixels and preserved without downscaling. For training, images were resized to 1920 × 1088 pixels to meet the padding requirements of multi-stage downsampling architectures while maintaining aspect ratio. Illumination conditions were categorized into three representative scenarios, namely cloudy, sunny, and scorching, each collected across multiple days to capture temporal variability in lighting and background complexity [13], [17]. Binary masks were manually annotated using dedicated labeling software under standardized annotation guidelines by two independent annotators. Annotation reliability was verified through stratified random sampling of 30 image–mask pairs, which achieved a Cohen’s Kappa of 0.9628, indicating very high consistency [18].

To improve robustness against environmental heterogeneity, online data augmentation was applied during training with randomized transformations in each batch [19]. Augmentation techniques were carefully selected to reflect realistic variations in the nursery environment. Brightness and contrast adjustments simulated natural illumination changes such as shading from clouds or intense sunlight [20]. CLAHE and histogram equalization enhanced local and global contrast to improve boundary visibility under uneven lighting [21]. Random gamma correction altered luminance distribution to mimic overexposure or underexposure, while HSV adjustments modified hue, saturation, and value to capture leaf coloration variability caused by lighting [22]. Geometric transformations such as horizontal and vertical flipping increased orientation diversity, reflecting different planting layouts and overlapping seedlings [23]. Random rotation and zoom were also applied to simulate varied leaf orientations and scale differences, reflecting changes in seedling position and camera distance in the nursery. Gaussian blur was introduced to reproduce motion blur or minor defocus, thereby improving resilience to imperfect capture conditions [24]. This augmentation strategy ensured that each epoch generated unique samples, reduced the risk of overfitting, and strengthened the model’s ability to generalize under diverse field conditions.

2.2 Dataset Summary

The dataset was specifically designed to capture the visual heterogeneity of open nursery environments where *Eucalyptus pellita* seedlings are cultivated. A total of 2,424 images were collected and systematically divided into three subsets: 1,939 images for training (80%), 242 images for validation (10%), and 243 images for testing (10%). The test set was stratified to ensure representative coverage of illumination scenarios, comprising 76 images under cloudy conditions, 63 images under sunny conditions, and 104 images under scorching conditions. This allocation guaranteed that model evaluation reflected realistic challenges posed by natural variability in lighting and background composition. Each lighting condition contributed unique visual challenges for segmentation. Cloudy conditions provided diffuse light that minimized shadows and highlights, clarifying leaf boundaries but reducing global contrast. Sunny conditions yielded relatively balanced luminance but introduced shadows from overlapping leaves and surrounding objects, often leading to false positives. Scorching conditions presented the most difficult scenario, as intense direct light created specular reflections that obscured surface textures and increased the likelihood of false negatives. Beyond illumination, additional variability was introduced by diverse nursery backgrounds such as concrete flooring, dry leaves, pot trays, irrigation pipes, and metallic racks. Leaf overlap was frequently observed, creating occlusions and complicating boundary detection, particularly at full resolution. By encompassing both controlled variability and realistic disturbances, the dataset established a robust benchmark for training and evaluation. It challenged models not only to detect leaf regions accurately under ideal conditions but also to generalize effectively across harsh lighting and cluttered backgrounds. This diversity makes the dataset a strong representation of operational conditions in industrial forestry nurseries, ensuring that segmentation models trained on it are applicable to real-world monitoring and management practices.

2.3 Model Architecture

This study employed three image segmentation architectures, with the Modified U-Net as the primary model. The other two models, SegNet and DeepLabv3+, were used solely for benchmarking so that the primary model’s performance could be evaluated objectively under identical training conditions. The Modified U-Net integrated a ResNet-50 encoder pre-trained on ImageNet, L2 regularization in the decoder, and a composite Binary Cross-Entropy and Dice loss. SegNet employed a VGG16 encoder pre-trained on ImageNet with an unpooling decoder based on pooling indices to preserve spatial details [24]. DeepLabv3+ utilized an Xception backbone with Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context and a decoder to refine spatial resolution [25]. All models were trained and evaluated under identical configurations to ensure fair comparison. A summary of the configurations is provided in Table 1.

Table 1. Configuration of the Architectures used in this Study

Parameter	Modified U-Net	SegNet	DeepLabv3+
Encoder	ResNet-50	VGG16	Xception
Pre-trained Weights	Yes (ImageNet)	Same as Modified U-Net	Same as Modified U-Net
Decoder	Transposed convolution + L2 regularization	Unpooling based on pooling indices	Transposed convolution (with ASPP before decoding)

Skip Connections	Yes	No	Not explicit
Input Size	1920 × 1088 pixels	Same as Modified U-Net	Same as Modified U-Net
Activation	ReLU; sigmoid output	Same as Modified U-Net	Same as Modified U-Net
Loss Function	Binary Cross-Entropy and Dice loss	Same as Modified U-Net	Same as Modified U-Net
Optimizer	Adam (initial lr 0.0001, ReduceLROnPlateau)	Same as Modified U-Net	Same as Modified U-Net
Batch Size	2	Same as Modified U-Net	Same as Modified U-Net
Maximum Epochs	1000 (with early stopping)	Same as Modified U-Net	Same as Modified U-Net
Data Augmentation	Horizontal/vertical flips, random rotation, zoom, blur, brightness/contrast adjustment	Same as Modified U-Net	Same as Modified U-Net
Model role	Primary	Baseline	Baseline

2.4 Evaluation and Metrics

Segmentation performance was comprehensively assessed using six metrics: Dice coefficient and Intersection over Union (IoU) as the primary measures of overlap quality, Precision and Recall as indicators of false positive control and detection completeness, F1 score as a balance between Precision and Recall, and Accuracy as a global measure of correctly classified pixels. Together, these six metrics capture shape conformity, edge preservation, and robustness under challenging illumination. The Dice coefficient (Equation 1) measures the spatial overlap between predicted and ground truth masks and is widely used as a segmentation metric because it balances false positives and false negatives [26]. The IoU (Equation 2) quantifies the ratio of intersection over union between predicted and ground truth regions, providing a stricter penalty for mismatches [27]. Precision (Equation 3) evaluates the proportion of correctly predicted leaf pixels relative to all predicted positives, while Recall (Equation 4) measures the proportion of correctly detected leaf pixels relative to all actual leaf pixels [28]. The F1 score (Equation 5) is the harmonic mean of Precision and Recall, emphasizing balance between the two [29]. Finally, Accuracy (Equation 6) captures the proportion of correctly classified pixels (both leaf and background) over the entire image, though it is less reliable under class imbalance where background dominates [30],

$$Dice\ Coefficient = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1}$$

$$IoU = \frac{A \cap B}{A \cup B} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

where,

TP: True Positive, correctly segmented leaf pixels
TN: True Negative, correctly identified background pixels
FP: False Positive, background pixels incorrectly classified as leaf
FN: False Negative, leaf pixels missed by segmentation

Dice coefficient and IoU were selected as the primary metrics because they directly evaluate segmentation overlap, which is the core objective of this task. Unlike Accuracy, which may be inflated in imbalanced datasets dominated by background pixels, Dice and IoU are more sensitive to errors at the object level [31]. Their complementary nature allows robust assessment: Dice emphasizes pixel-wise agreement, while IoU imposes stricter penalties for disagreement. Benchmarking was conducted as a controlled comparative study of the Modified U-Net, SegNet, and DeepLabv3+. The test set consisted of 243 images (10% of the dataset), stratified into 76 cloudy, 63 sunny, and 104 scorching samples. Inference was performed at the native resolution of 1920 × 1080 with a binary threshold of 0.5 and no post-processing. All metrics were computed on a per-image basis and summarized using the median and interquartile range (IQR). Error rates were further normalized to the number of leaf pixels in the ground truth to enable fair comparison across conditions.

For qualitative evaluation, representative panels were generated for each lighting condition, including the original image, the ground truth mask, and segmentation outputs from the three models. Semi-transparent overlays were applied with green indicating TP, uncolored background for TN, red for FP, and blue for FN. This facilitated visual inspection of shape conformity, boundary preservation, and detection of small leaves, as well as error sources such as glare and shadow. To assess statistical significance of performance differences, the Kruskal–Wallis statistic was calculated using Equation 7.

$$H = \frac{12}{N(N+1)} \sum \frac{R_j^2}{n_j} - (N+1) \tag{7}$$

Where *N* is the total number of observations, *Rj* the sum of ranks for group *j*, and *nj* the sample size of group *j*. According to Tomczak and Tomczak [33], ε^2 values can be interpreted as negligible (<0.01), small (0.01–0.08), medium (0.08–0.26), and large (>0.26).

3. Results and Discussion

This section presents both quantitative and qualitative results for the three segmentation models (Modified U-Net, SegNet, and DeepLabv3+) evaluated under cloudy, sunny, and scorching illumination using the benchmarking protocol described in Section 2.4. Tables are used to report numerical performance across all evaluation metrics, while figures provide complementary qualitative comparisons that illustrate segmentation outcomes and typical error patterns under varying lighting conditions.

3.1 Quantitative Evaluation

The quantitative results in Table 2 consistently demonstrate the superiority of the Modified U-Net over the baseline models, with performance variations across lighting conditions that align with the qualitative observations.

Table 2. Segmentation Performance (median [IQR]) of Three Models under Different Lighting Conditions

Model	Metric	Cloudy	Sunny	Scorching
Modified U-Net	Dice coefficient	0.872 [0.479]	0.841 [0.390]	0.854 [0.509]
	IoU	0.773 [0.634]	0.725 [0.538]	0.745 [0.659]
	Precision	0.952 [0.612]	0.841 [0.518]	0.907 [0.638]
	Recall	0.862 [0.191]	0.890 [0.162]	0.882 [0.166]
	F1 score	0.872 [0.479]	0.841 [0.390]	0.854 [0.509]
	Accuracy	0.816 [0.259]	0.873 [0.247]	0.886 [0.232]
SegNet	Dice coefficient	0.538 [0.454]	0.533 [0.385]	0.476 [0.479]
	IoU	0.368 [0.513]	0.363 [0.438]	0.312 [0.541]
	Precision	0.924 [0.088]	0.953 [0.082]	0.947 [0.081]
	Recall	0.379 [0.572]	0.377 [0.458]	0.318 [0.570]
	F1 score	0.538 [0.454]	0.533 [0.385]	0.476 [0.479]
	Accuracy	0.713 [0.369]	0.870 [0.213]	0.742 [0.284]
DeepLabv3+	Dice coefficient	0.648 [0.448]	0.529 [0.415]	0.467 [0.429]
	IoU	0.479 [0.458]	0.359 [0.458]	0.305 [0.460]
	Precision	0.918 [0.089]	0.924 [0.109]	0.883 [0.079]
	Recall	0.534 [0.513]	0.389 [0.454]	0.316 [0.532]
	F1 score	0.648 [0.448]	0.529 [0.415]	0.467 [0.429]
	Accuracy	0.764 [0.387]	0.880 [0.223]	0.733 [0.267]

Dice coefficient serves as the primary indicator of overlap between predicted segmentations and the ground truth. The Modified U-Net achieved the highest Dice under cloudy conditions (0.872), decreased under sunny illumination (0.841), and remained strong under scorching illumination (0.854). The decline under sunny conditions reflects the effect of glare-induced False Positives, while the recovery under scorching conditions suggests relative resilience to extreme illumination variability. In contrast, SegNet (0.476–0.538) and DeepLabv3+ (0.467–0.648) consistently scored lower, indicating systematic under-segmentation and limited generalization in natural environments. Intersection over Union (IoU) reinforces these findings. The Modified U-Net yielded its highest IoU of 0.773 under cloudy conditions, dropped to 0.725 under sunny illumination, and recovered to 0.745 under scorching conditions, reflecting overall segmentation stability. SegNet (0.312–0.368) and DeepLabv3+ (0.305–0.479) produced substantially lower IoU values, with wider variability, highlighting their inconsistent coverage of leaf regions.

Precision reflects the proportion of correctly classified positive pixels. The Modified U-Net maintained high Precision across all conditions (0.841–0.952), peaking under cloudy illumination. The lowest Precision occurred under sunny conditions (0.841), consistent with glare artifacts that produced localized False Positives. Although these precision drops under sunny illumination reflect glare-induced FP, the overall Dice/IoU distributions remained statistically comparable across lighting (see Section 3.2). Interestingly, SegNet (0.924–0.953) and DeepLabv3+ (0.883–0.924) also achieved high Precision, but this was offset by extremely poor Recall, resulting in weak overall segmentation performance. Recall measures the sensitivity of models in detecting all leaf regions. The Modified U-Net achieved stable Recall values across all lighting conditions (0.862–0.890), indicating consistent detection of the majority of leaves, even under adverse illumination. In contrast, SegNet (0.318–0.379) and DeepLabv3+ (0.316–0.534) consistently showed low Recall, underscoring their tendency toward False Negatives and explaining the under-segmentation observed in qualitative overlays. F1 score, the harmonic mean of Precision and Recall, highlights the balance between accuracy and sensitivity. The Modified U-Net achieved its highest F1 score under cloudy conditions (0.872), its lowest under sunny illumination (0.841), and maintained stable performance under scorching illumination (0.854). By comparison, SegNet (0.476–0.538) and DeepLabv3+ (0.467–0.648) fell significantly behind, reinforcing their inability to balance Precision and Recall.

Accuracy, defined as the proportion of correctly classified pixels, exhibited a slightly different trend. For the Modified U-Net, Accuracy was relatively high across all conditions (0.816–0.886), with increases under sunny (0.873) and scorching illumination (0.886). This trend can be explained by the dominance of background pixels; even though Precision decreased due to glare-induced False Positives, large non-leaf regions were still correctly classified as negatives, sustaining overall Accuracy. For SegNet and DeepLabv3+, Accuracy (0.713–0.880) fluctuated more widely and did not align with Dice or IoU, suggesting that this metric is less informative for class-imbalanced segmentation tasks. The Modified U-Net demonstrated robust segmentation across all metrics, with particularly strong Dice, IoU, Recall, and F1 scores. The decrease in Precision observed under sunny conditions highlights a specific vulnerability to glare, which could potentially be mitigated through glare-aware augmentation strategies. In contrast, SegNet and DeepLabv3+ maintained high Precision but suffered from critically low Recall, reflecting an overly conservative segmentation behavior that reduced False Positives at the expense of missing large portions of leaf regions. This tendency limits their reliability for field applications that require comprehensive detection. Accuracy, on the other hand, must be interpreted with caution, as its apparent improvement under certain conditions is largely attributable to background dominance rather than true segmentation performance. Consequently, metrics such as Dice, IoU, and F1 provide a more reliable and representative evaluation of segmentation quality.

3.2 Statistical Analysis

Non-parametric Kruskal–Wallis tests [32], were conducted on per-image Dice coefficient and IoU values, with epsilon-squared (ε^2) reported as an effect size [33]. In this context, the p-value assesses the null hypothesis that median segmentation performance does not differ across lighting conditions ($p < 0.05$ indicates a statistically significant difference), while ε^2 quantifies the proportion of variance explained by the factor, interpreted as negligible (<0.01), small (0.01–0.08), medium (0.08–0.26), and large (>0.26). All analyses were performed on the independent test set comprising 243 images (76 cloudy, 63 sunny, 104 scorching). For the Modified U-Net, the Kruskal–Wallis test yielded $H = 4.012$, $df = 2$, $p = 0.1345$, with an effect size of $\varepsilon^2 = 0.008$, indicating a negligible influence of illumination on segmentation outcomes. Similarly, SegNet ($H = 2.915$, $p = 0.2328$, $\varepsilon^2 = 0.004$) and DeepLabv3+ ($H = 1.248$, $p = 0.5357$, $\varepsilon^2 = -0.003$) showed no significant differences, with effect sizes approaching zero. The detailed outcomes of the Kruskal–Wallis tests for all models and metrics are summarized in Table 3.

Table 3. Kruskal–Wallis Tests on per-image Dice and IoU across Lighting Conditions (Cloudy, Sunny, Scorching)

Model	Metric	H (df=2)	p-value	ε²	Notes
Modified U-Net	Dice	4.012	0.1345	0.008	Not significant
Modified U-Net	IoU	4.012	0.1345	0.008	Not significant
SegNet	Dice	2.915	0.2328	0.004	Not significant
SegNet	IoU	2.915	0.2328	0.004	Not significant
DeepLabv3+	Dice	1.248	0.5357	−0.003	Not significant
DeepLabv3+	IoU	1.248	0.5357	−0.003	Not significant

Table 3 shows that none of the models exhibited statistically significant differences across lighting conditions (all p > 0.05). These findings demonstrate that segmentation performance, as measured by Dice and IoU, was statistically comparable under cloudy, sunny, and scorching conditions for all models. In practical terms, this indicates that the Modified U-Net maintained robust performance across natural lighting variability. While localized errors such as glare-induced false positives in sunny scenes and shadow-related artifacts under scorching illumination reduced precision in certain cases, they did not alter the overall overlap distributions sufficiently to produce statistically significant changes. In contrast, SegNet and DeepLabv3+ yielded consistently lower Dice and IoU values regardless of illumination, reflecting limited adaptability rather than robustness. This reinforces the importance of illumination-aware augmentation strategies to further mitigate localized errors despite the overall stability of the Modified U-Net.

3.3 Qualitative Evaluation

Figure 2 presents representative qualitative overlays of segmentation results under cloudy illumination. Panel (a) shows the original input image, while panels (b)-(d) display segmentation outputs from different models with color-coded overlays. The Modified U-Net (Figure 2b) achieves the most accurate segmentation, with extensive green overlays covering nearly all leaf regions and only minor red or blue artifacts. In contrast, SegNet (Figure 2c) produces more red overlays along the background, particularly over textured non-leaf surfaces, and leaf boundaries appear less well-defined. DeepLabv3+ (Figure 2d) shows more pronounced under-segmentation, with large blue overlays indicating missed leaf regions, especially in overlapping foliage and shaded areas. Error analysis confirms that under cloudy conditions the most common errors were small false positives along background clutter and false negatives at thin leaf margins. These localized artifacts did not alter the overall Dice/IoU distributions sufficiently to yield statistical significance, consistent with the results in Section 3.2. Quantitative error analysis under cloudy illumination further supports these observations. The Modified U-Net achieved an average FP rate of 0.061 and FN rate of 0.118, showing a strong balance between sensitivity and specificity. SegNet maintained a low FP rate of 0.043 but exhibited a very high FN rate of 0.621, indicating systematic under-segmentation. DeepLabv3+ presented intermediate values, with FP = 0.071 and FN = 0.457. These findings confirm that cloudy conditions provide the most favorable imaging scenario, allowing the Modified U-Net to achieve the most reliable segmentation performance.

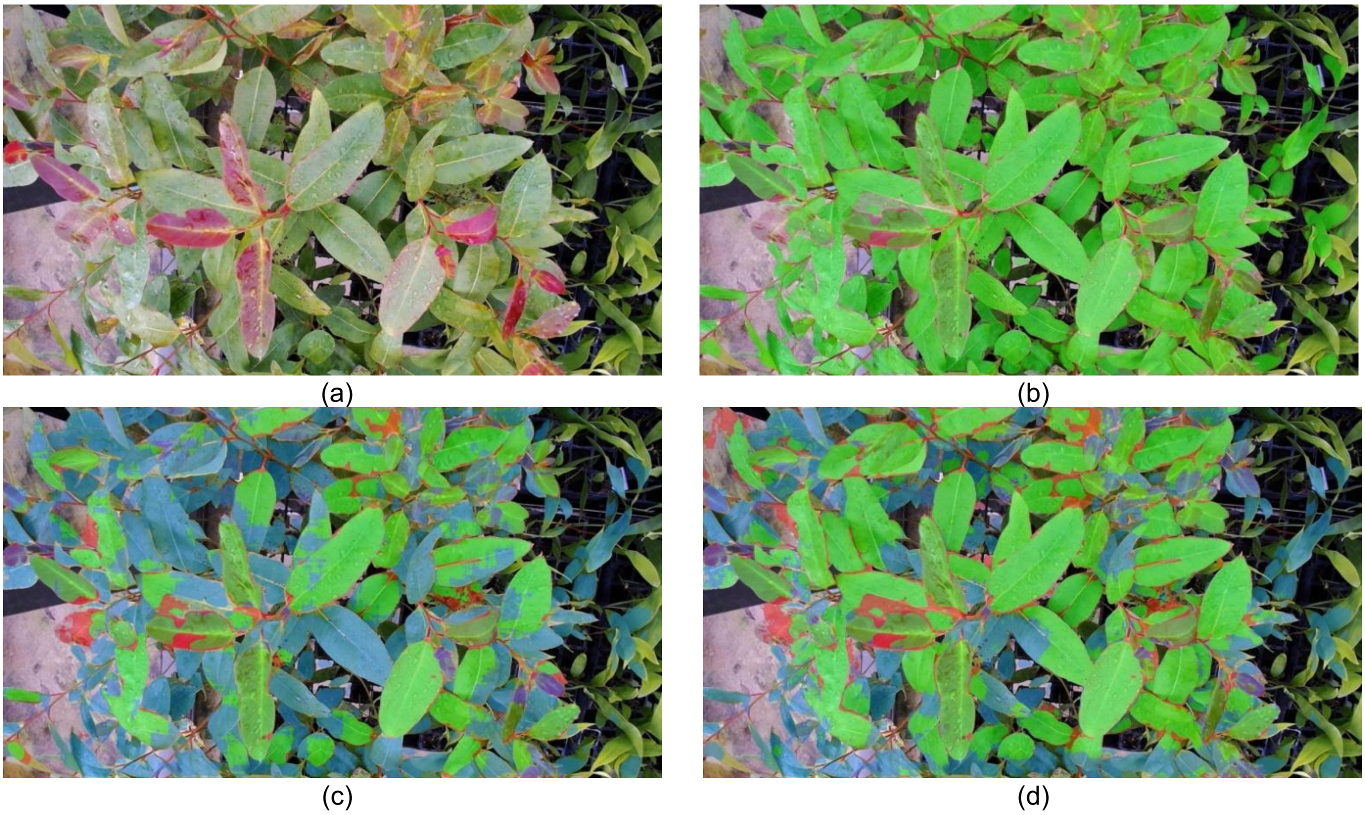

Figure 2. Qualitative Comparison of Segmentation Outputs under Cloudy Conditions: (a) Original Image, (b) Modified U-Net Segmentation, (c) SegNet Segmentation, and (d) DeepLabv3+ Segmentation

Figure 3 compares segmentation results under sunny illumination. The Modified U-Net (Figure 3b) maintains strong segmentation performance, with dominant green overlays representing True Positive regions. However, glare on certain leaf surfaces introduced localized False Positives, where bright specular reflections were mistakenly classified as leaf areas. SegNet (Figure 3c) exhibits more extensive False Positives in shadowed background regions, and its boundary precision remains weaker compared to the Modified U-Net. DeepLabv3+ (Figure 3d) again demonstrates under-segmentation, with widespread blue overlays indicating False Negatives in regions affected by specular highlights. Error analysis indicates that sunny illumination posed the greatest qualitative challenge due to glare and directional shadows. These glare-induced false positives explain localized drops in precision but did not produce statistically significant shifts in Dice/IoU distributions.

The quantitative FP/FN analysis confirms this trend. The Modified U-Net recorded an FP rate of 0.079 and FN rate of 0.162, showing that glare primarily increased false positives while recall remained relatively stable. SegNet again produced a very low FP rate (0.049) but extremely high FN (0.664), confirming its conservative prediction bias. DeepLabv3+ demonstrated FP = 0.087 and FN = 0.552, highlighting its difficulty in capturing leaf regions affected by strong reflections. These values align with the qualitative finding that sunny illumination is the most error-prone scenario.

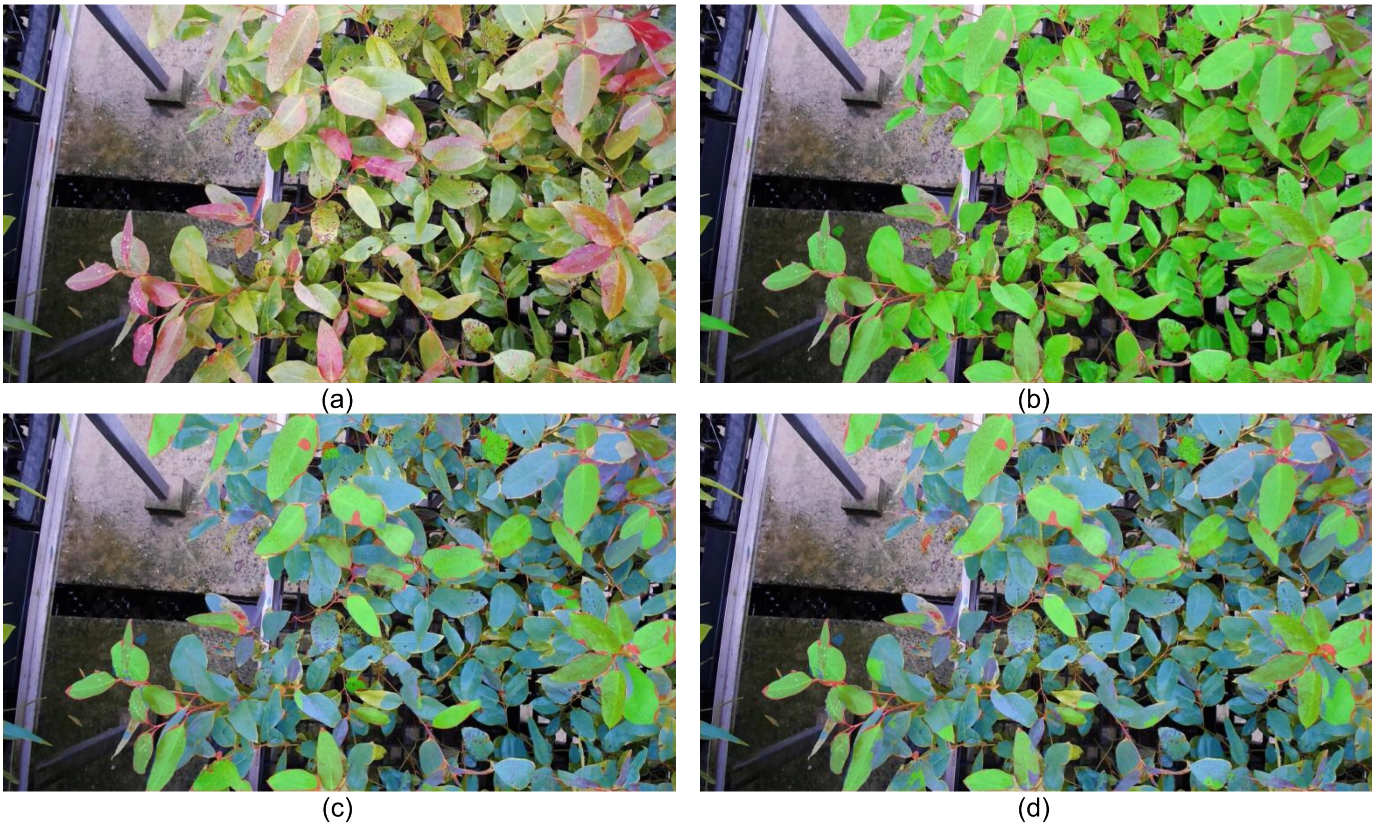

Figure 3. Qualitative Comparison of Segmentation Outputs under Sunny Conditions: (a) Original Image, (b) Modified U-Net Segmentation, (c) SegNet Segmentation, and (d) DeepLabv3+ Segmentation

Figure 4 illustrates segmentation performance under scorching illumination. The Modified U-Net (Figure 4b) continues to outperform the baseline models, preserving the majority of leaf regions with extensive green overlays indicating True Positive coverage, although shadow-dense areas introduced moderate False Positives. SegNet (Figure 4c) generates a large number of False Positives in dark background regions, particularly along high-contrast edges between leaves and soil. DeepLabv3+ (Figure 4d) again exhibits substantial False Negatives, with extensive blue overlays highlighting missed foliage. Error analysis under scorching conditions reveals that intense direct sunlight introduced both shadow-induced false positives and glare-induced false negatives. These shadow-related errors contributed to precision fluctuations but were insufficient to shift Dice/IoU distributions to statistical significance.

Quantitative results provide further evidence of this pattern. The Modified U-Net showed FP = 0.072 and FN = 0.177, slightly higher FN than under cloudy or sunny conditions but still balanced overall. SegNet recorded FP = 0.051 with FN = 0.642, again reflecting massive under-segmentation. DeepLabv3+ produced FP = 0.091 and FN = 0.489, demonstrating persistent under-detection of leaf regions under intense illumination. These values confirm that scorching conditions primarily elevate false negatives, particularly in SegNet and DeepLabv3+, while the Modified U-Net maintains comparatively stable performance.

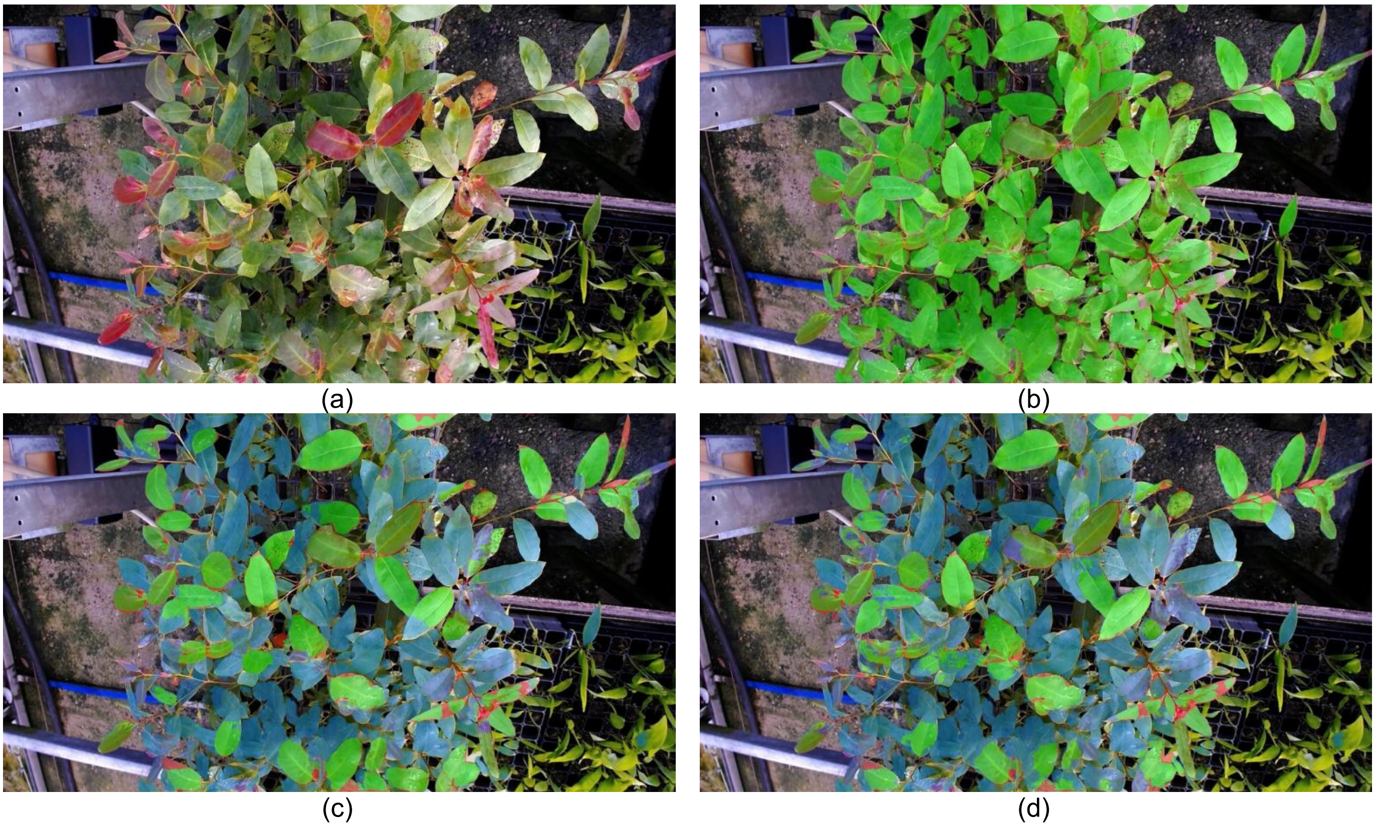

Figure 4. Qualitative Comparison of Segmentation Ooutputs under Scorching Conditions: (a) Original Image, (b) Modified U-Net Segmentation, (c) SegNet Segmentation, and (d) DeepLabv3+ Segmentation

3.4 State-of-the-Art Comparison

To contextualize the above quantitative and qualitative findings, we compare our overlap scores with recent leaf-segmentation studies. In field orchard imagery, an Eff-UNet–based apple-leaf pipeline reported an average test IoU of 0.72 [34], indicating feasible but still modest pixel-wise overlap under real-world variability. On a related leaf-image task, an improved UNet (RS-UNet) achieved mIoU = 0.798 [35]. In comparison, our Modified U-Net attains IoU = 0.773 under cloudy scenes, placing it above the orchard baseline and approaching the improved UNet result obtained on a different dataset, while being validated under natural, variable illumination (cloudy/sunny/scorching). These cross-study contrasts reinforce that the proposed architecture delivers competitive overlap quality for leaf segmentation in outdoor nursery conditions, acknowledging that absolute comparability is limited by dataset differences and the use of IoU versus mIoU.

3.5 Discussion and Implications

The Modified U-Net consistently demonstrated accurate and stable segmentation under natural nursery conditions, outperforming SegNet and DeepLabv3+ across all lighting scenarios. Overall robustness was confirmed by Kruskal–Wallis tests on per-image Dice and IoU, which detected no significant differences across cloudy, sunny, and scorching conditions (all $p > 0.05$). This indicates that the model maintained statistically comparable performance distributions despite variations in illumination. Nevertheless, qualitative error analysis highlighted systematic patterns that explain fluctuations in certain metrics. Sunny scenes remained the most challenging due to glare-induced false positives, which reduced precision while recall stayed high. Scorching conditions introduced shadow-related false positives, but their influence was less pronounced than glare, resulting in localized errors without altering global Dice or IoU distributions. In contrast, cloudy conditions provided diffuse illumination that minimized both glare and shadows, enabling more consistent segmentation with fewer artifacts. These observations suggest that while the Modified U-Net preserved stable performance across lighting conditions in statistical terms, illumination still shaped error modes that have practical relevance. For operational deployment, cloudy acquisition remains preferable whenever possible, whereas glare- and shadow-aware augmentation strategies should be incorporated into training pipelines to mitigate localized failures in sunny and scorching scenarios. Such adaptations would further enhance precision without compromising the robustness already demonstrated in global metrics.

4. Conclusion

This study evaluated the performance of Modified U-Net, SegNet, and DeepLabv3+ for leaf segmentation of *Eucalyptus pellita* seedlings under three natural illumination conditions (cloudy, sunny, and scorching) in open nursery environments. The Modified U-Net, incorporating a ResNet-50 encoder, L2 regularization, and a composite Binary Cross-Entropy and Dice loss, consistently outperformed the baseline models across all scenarios. On the independent test set (243 images; 76 cloudy, 63 sunny, 104 scorching), Kruskal–Wallis tests on per-image Dice and IoU found no significant differences across lighting conditions ($p > 0.05$), indicating stable segmentation performance under natural illumination variability. Qualitative overlays and FP/FN analyses nonetheless revealed localized challenges: glare-induced false positives in sunny scenes and shadow-related artifacts under scorching light. These error modes explain precision fluctuations while leaving overall Dice and IoU distributions statistically comparable. Taken together, the findings demonstrate that the Modified U-Net delivers robust segmentation in realistic nursery conditions, with consistent performance across diverse illumination scenarios. The main contributions of this study are threefold: (i) the integration of ResNet-50 encoder, L2 regularization, and BCE+Dice loss as a novel architectural modification for outdoor segmentation; (ii) empirical validation on a large, annotated dataset with high inter-annotator agreement ($\kappa = 0.9628$); and (iii) demonstration of stable performance across challenging natural lighting. For practical deployment, glare- and shadow-aware augmentation strategies and computational optimizations (e.g., quantization, lightweight encoders) are recommended to mitigate localized errors and enable near real-time inference on edge devices.

References

[1] K. von Rintelen, E. Arida, and C. Häuser, “A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries,” *Res. Ideas Outcomes*, vol. 3, 2017. https://doi.org/10.3897/rio.3.e20860

[2] BPS, “Statistik Produksi Kehutanan 2023,” *Badan Pus. Stat.*, p. 32, 2023.

[3] Peraturan Presiden No.66 Tahun 2017 tentang Koordinasi Startegis Lintas sektoral Penyelengaraan Pelayanan Kepemudaan, “Lembaran Negara Republik,” *Rencana Umum Energi Nas.*, no. 73, pp. 1–6, 2017.

[4] M. A. Inail, E. B. Hardiyanto, and D. S. Mendham, “Growth responses of Eucalyptus pellita F. Muell plantations in south sumatra to macronutrient fertilisers following several rotations of Acacia mangium willd,” 2019. https://doi.org/10.3390/F10121054

[5] S. C. Grossnickle, “Seedling establishment on a forest restoration site-An ecophysiological perspective,” *Reforesta*, vol. 6, pp. 110–139, 2018. https://doi.org/10.21750/REFOR.6.09.62

[6] S. Jayathunga, G. D. Pearse, and M. S. Watt, “Unsupervised Methodology for Large-Scale Tree Seedling Mapping in Diverse Forestry Settings Using UAV-Based RGB Imagery,” 2023. https://doi.org/10.3390/rs15225276

[7] G. Papadopoulos, S. Arduini, H. Uyar, V. Psiroukis, A. Kasimati, and S. Fountas, “Economic and environmental benefits of digital agricultural technologies in crop production: A review,” *Smart Agric. Technol.*, vol. 8, p. 100441, 2024. https://doi.org/10.1016/j.atech.2024.100441

[8] Y. Diez, S. Kentsch, M. Fukuda, M. L. L. Caceres, K. Moritake, and M. Cabezas, “Deep learning in forestry using uav-acquired rgb data: A practical review,” 2021. https://doi.org/10.3390/rs13142837

[9] Z. Luo, W. Yang, Y. Yuan, R. Gou, and X. Li, “Semantic segmentation of agricultural images: A survey,” *Inf. Process. Agric.*, vol. 11, no. 2, pp. 172–186, 2024. https://doi.org/10.1016/j.inpa.2023.02.001

[10] K. Li, L. Zhang, B. Li, S. Li, and J. Ma, “Attention-optimized DeepLab V3 + for automatic estimation of cucumber disease severity,” *Plant Methods*, vol. 18, no. 1, p. 109, 2022. https://doi.org/10.1186/s13007-022-00941-8

[11] R. Ma *et al.*, “Local refinement mechanism for improved plant leaf segmentation in cluttered backgrounds,” *Front. Plant Sci.*, vol. Volume 14, 2023. https://doi.org/10.3389/fpls.2023.1211075

[12] K. Yang, W. Zhong, and F. Li, “Leaf Segmentation and Classification with a Complicated Background Using Deep Learning,” *Agronomy*, vol. 10, p. 1721, Nov. 2020. https://doi.org/10.3390/agronomy10111721

[13] A. Silwal, T. Parhar, F. Yandun, H. Baweja, and G. Kantor, “A Robust Illumination-Invariant Camera System for Agricultural Applications,” *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 3292–3298, 2021. https://doi.org/10.1109/IROS51168.2021.9636542

[14] J. Mitchel, T. Gao, E. Cole, V. Petukhov, and P. V. Kharchenko, “Impact of Segmentation Errors in Analysis of Spatial Transcriptomics Data.,” *bioRxiv*, p. 2025.01.02.631135, Jan. 2025. https://doi.org/10.1101/2025.01.02.631135

[15] F. J. Hutapea, C. J. Weston, D. Mendham, and L. Volkova, “Sustainable management of Eucalyptus pellita plantations: A review,” *For. Ecol. Manage.*, vol. 537, p. 120941, 2023. https://doi.org/10.1016/j.foreco.2023.120941

[16] M. N. Megat Mohamed Nazir, R. Terhem, A. R. Norhisham, S. Mohd Razali, and R. Meder, “Early monitoring of health status of plantation-grown eucalyptus pellita at large spatial scale via visible spectrum imaging of canopy foliage using unmanned aerial vehicles,” 2021. doi: 10.3390/f12101393. https://doi.org/10.3390/f12101393

[17] Y. Wang, Z. Yang, K. Gert, and H. A. Khan, “The impact of variable illumination on vegetation indices and evaluation of illumination correction methods on chlorophyll content estimation using UAV imagery,” *Plant Methods*, vol. 19, no. 1, p. 51, 2023. https://doi.org/10.1186/s13007-023-01028-8

[18] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Sep. 1977. https://doi.org/10.2307/2529310

[19] X. Song *et al.*, “Agricultural Image Processing: Challenges, Advances, and Future Trends,” 2025. https://doi.org/10.3390/app15169206

[20] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, p. 60, 2019. https://doi.org/10.1186/s40537-019-0197-0

[21] Samsuzzaman *et al.*, “Automated Seedling Contour Determination and Segmentation Using Support Vector Machine and Image Features,” 2024. https://doi.org/10.3390/agronomy14122940

[22] W. Zhou, A. Zyner, S. Worrall, and E. Nebot, “Adapting Semantic Segmentation Models for Changes in Illumination and Camera Perspective,” *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 461–468, 2019. https://doi.org/10.1109/lra.2019.2891027

[23] R. Zenkl *et al.*, “Outdoor Plant Segmentation With Deep Learning for High-Throughput Field Phenotyping on a Diverse Wheat Dataset,” *Front. Plant Sci.*, vol. 12, 2022. https://doi.org/10.3389/fpls.2021.774068

[24] K. Alomar, H. I. Aysel, and X. Cai, “Data Augmentation in Classification and Segmentation: A Survey and New Strategies,” 2023. https://doi.org/10.3390/jimaging9020046

[25] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-01234-2_49

[26] F. Milletari, N. Navab, and S. A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, 2016. https://doi.org/10.1109/3DV.2016.79

[27] P. Jaccard, “the Distribution of the Flora in the Alpine Zone.,” *New Phytol.*, vol. 11, no. 2, pp. 37–50, Feb. 1912. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. https://doi.org/10.1017/cbo9780511809071

[29] S. Sturges and M. Brown, “Polypsychopharmacy,” *Bull. Menninger Clin.*, vol. 39, no. 3, pp. 274–279, 1975. https://doi.org/10.1097/01.jcp.0000177847.77791.99

[30] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009. https://doi.org/10.1016/j.ipm.2009.03.002

[31] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Med. Imaging*, vol. 15, no. 1, p. 29, 2015. https://doi.org/10.1186/s12880-015-0068-x

[32] W. H. Kruskal and W. A. Wallis, “Use of Ranks in One-Criterion Variance Analysis,” *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, Dec. 1952. https://doi.org/10.1080/01621459.1952.10483441

[33] Tomczak M and Tomczak E, “The need to report effect size estimates revisited. An overview of some recommended measures of effect size,” *Trends Sport Sci.*, vol. 21, no. 1, pp. 19–25, Jan. 2014.

[34] A. Uryasheva, A. Kalashnikova, D. Shadrin, K. Evteeva, E. Moskovtsev, and N. Rodichenko, “Computer vision-based platform for apple leaves segmentation in field conditions to support digital phenotyping,” *Comput. Electron. Agric.*, vol. 201, no. September 2021. https://doi.org/10.1016/j.compag.2022.107269

[35] J. Fu, Y. Zhao, and G. Wu, “Potato Leaf Disease Segmentation Method Based on Improved UNet,” *Appl. Sci.*, vol. 13, no. 20, 2023. https://doi.org/10.3390/app132011179