Semantic Segmentation Comparison of *Prata Catarina* Banana Bunches Using Convolutional Neural Network Models

Emannuel Diego G. de Freitas[1,2], José Herbert R. Moreno[2], Yuri Costa G. da Silva[2], Erica Lima Silva[2], Danielo G. Gomes[1]

[1]Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, Fortaleza-CE, Brasil.
[2]Instituto Federal de Educação Ciência e Tecnologia do Ceará, campus Iguatu, Iguatu-CE, Brasil.
[3]Image Pesquisas, Av. Humberto Monte, Bloco 310, Galpão 17, Fortaleza-CE, Brasil.

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

The identification and classification of banana ripening stages are essential for production assessment in large scale plantations, enabling efficient harvest monitoring and ensuring fruit quality for commercialization. This study presents a comparative evaluation of three deep learning architectures applied to the semantic segmentation of Prata Catarina banana bunches, aiming to support automated monitoring systems and decision-making tools for precision agriculture applications. The evaluated models were trained and tested under standardized experimental conditions using a dataset composed of 1,017 images of banana plants, manually annotated for the segmentation of bunches and rachises and managed through the Roboflow platform. Image acquisition was performed using multiple smartphone devices, resulting in varied resolutions and real-field acquisition conditions that reflect practical agricultural scenarios. All models were implemented in Python using TensorFlow and Keras frameworks, adopting unified hyperparameters to ensure a fair and consistent comparison. A hybrid BCE–Jaccard Loss function was employed to optimize pixel-level classification performance and segmentation similarity between predictions and reference masks. Model performance was evaluated using Precision, Recall, Dice Score, and Loss, enabling a comprehensive quantitative assessment of segmentation accuracy and generalization capability. Experimental results indicated that U-Net achieved the best overall performance during validation, demonstrating superior ability to preserve spatial information and accurately delineate banana bunch contours. PSP-Net showed balanced performance and the highest recall, indicating strong sensitivity in identifying relevant regions, although with a slight increase in false positives. In contrast, DeepLabV3+ achieved high training performance but exhibited limited generalization during validation, suggesting overfitting under the adopted training conditions and dataset size limitations. The results emphasize that architectural selection should consider dataset scale, computational resources, and task complexity. Overall, the findings demonstrate that semantic segmentation based on deep learning represents a promising technological approach for banana crop monitoring, contributing to the development of embedded systems capable of automated detection, data collection, and intelligent decision support in precision agriculture.

Keywords

Banana farming; Computer vision; Automated harvest monitoring; U-Net benchmarking.

Introduction

Due to its high nutritional value and versatility, the banana stands as one of the most widely produced fruits and maintains strong demand in the international market, playing a vital socioeconomic role for producing countries (FAO, 2024). In Brazil, this relevance is supported not only by its economic impact but also by the country's position as the world's largest consumer of the fruit, reinforcing the need to optimize production processes (EMBRAPA, 2025).

Despite the sector's importance, processes such as bunch counting and harvest assessment still rely on manual and subjective evaluations. This practice creates efficiency bottlenecks and makes production vulnerable to human error and high labor demand (Baglat et al., 2023). In this context, automation through computer vision emerges as an essential solution to standardize monitoring practices and enhance technological competitiveness in banana cultivation.

The adoption of computer vision techniques, particularly semantic segmentation, is justified by the need to overcome the limitations of traditional methods in complex agricultural environments. As highlighted by He et al. (2023), deep learning approaches enable the extraction of detailed morphological features even in scenarios with low chromatic contrast—such as green targets against similarly colored backgrounds—thereby ensuring greater robustness, precision, and stability for automated field monitoring.

Recent studies have demonstrated the effectiveness of Deep Learning and Neural Network methods as viable alternatives for automating processes within the banana production chain. Aradhana et al. (2024) applied U-Net for fruit freshness identification through pixel-wise segmentation, achieving a Mean Intersection over Union (MIoU) of 94.9% in multi-fruit scenarios, highlighting the model's robustness under contextual variability. Similarly, Elinisa et al. (2024) demonstrated the effectiveness of U-Net in segmenting diseases such as Fusarium Wilt and Black Sigatoka, achieving a Dice Coefficient of 96.45% and an IoU of 93.23% in accurately delineating affected regions.

Multiscale architectures designed for semantic segmentation have also been applied directly to banana bunch structures. He et al. (2023) proposed an improved DeepLabv3 architecture for banana crown segmentation, achieving an MIoU of 85.75% with a significant reduction in parameters, indicating suitability for embedded applications. Likewise, Wu et al. (2023) employed DeepLabv3+ for segmentation and counting of banana hands across different phenological stages, obtaining an MIoU of 87.8% and a final accuracy of 93.2%. Despite these advances, existing studies often focus on specific structures or require additional processing stages, and comparative analyses among architectures applied to the complete banana bunch under real cultivation conditions remain limited.

Similarly, PSPNet has demonstrated consistent performance in crop and fruit delineation tasks. García-Downing et al. (2025) applied PSPNet with ResNet-34 for identifying banana and plantain cultivation areas in satellite imagery, achieving a Dice score of 76.2% and accuracy of 93.6%. Complementarily, Lu et al. (2022) reported competitive PSPNet performance in fruit segmentation within complex environments, showing strong IoU results and highlighting its ability to capture global contextual information in agricultural scenarios.

Therefore, although significant advances have been achieved using architectures such as U-Net, DeepLabv3+, and PSPNet in agriculture, gaps remain regarding direct comparisons of different models specifically applied to full banana bunch segmentation under standardized conditions. In this context, evaluating these architectures under identical experimental settings and evaluation criteria becomes essential to identify the most suitable approach for supporting banana cultivation

systems. Thus, this work proposes a comparative analysis of deep neural network architectures applied to the semantic segmentation of banana bunches, aiming to contribute to the advancement of automated decision-support systems within the banana production chain.

Convolutional Neural Network Architectures

U-Net Architecture

U-Net is a convolutional neural network architecture designed for fast and efficient image segmentation, improving upon traditional sliding-window convolutional networks through the use of innovative techniques and methods. Unlike earlier models that relied solely on contracting paths, U-Net features a distinctive U-shaped architecture composed of both contracting and expansive paths.

Input images and their corresponding segmentation maps are used to train the network using stochastic gradient descent implemented in Caffe (Jia et al., 2014). Due to the use of unpadded convolutions, the output image becomes smaller than the input by a constant border width, which minimizes computational overhead and maximizes GPU memory utilization. Consequently, it is preferable to use larger input patches rather than large batch sizes, often reducing the batch size to a single image (Ronneberger et al., 2015).

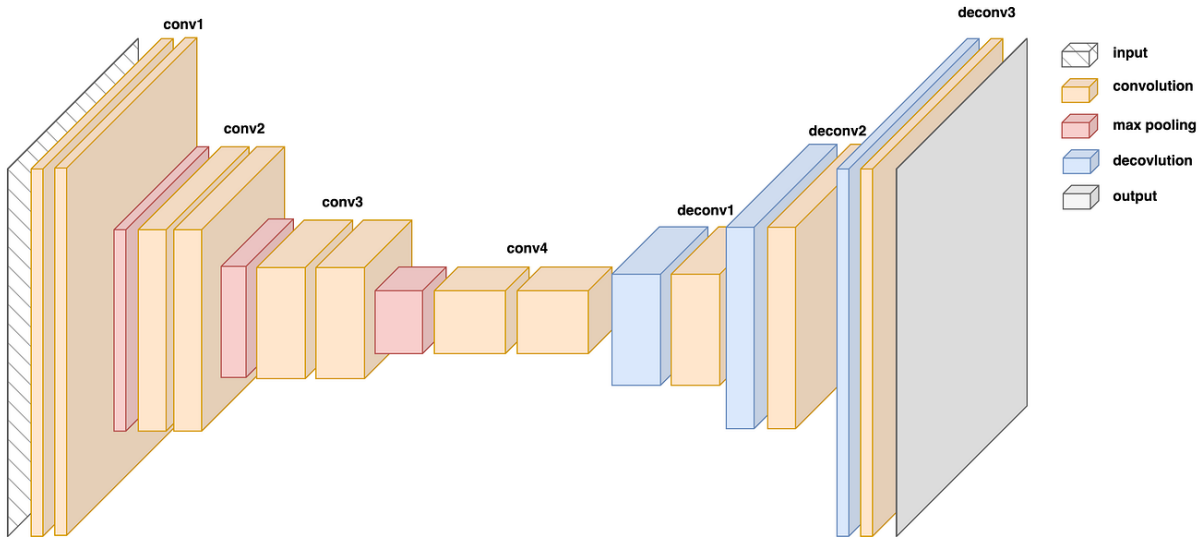


Figure 1. U-Net Architecture Diagram

As illustrated in Figure 1, the architecture consists of a contracting path (left side) and an expansive path (right side). The left side follows a typical convolutional neural network architecture, consisting of repeated applications of two 3 × 3 unpadded convolutions, each followed by a Rectified Linear Unit (ReLU) activation and a 2 × 2 max-pooling operation with a stride of 2 for downsampling. At each downsampling step, the number of feature channels is doubled (Ronneberger et al., 2015).

PSP-Net Architecture

The PSP-Net (Pyramid Scene Parsing Network) architecture, developed by Hengshuang et al. (2017), was designed to overcome the limitations of traditional models in capturing global contextual information in complex images. The main feature of this network is its Pyramid Pooling Module, which enables feature fusion across four different pyramid scales. These scales extract intrinsic features from the input at multiple resolutions (pyramidal scales), aggregating information from subregions of varying sizes. By incorporating multi-scale contextual information, PSPNet is able to better understand the relationship between the object of interest and its surrounding environment, thereby reducing classification errors in isolated pixels.

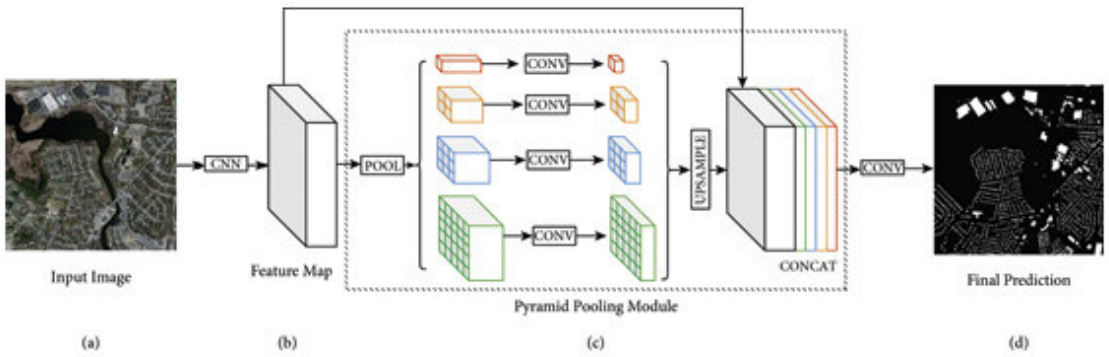


Figure 2. PSP-Net Architecture Diagram

Using Figure 2 as a reference, the processing begins with the generation of feature maps from the input image, in which contextual information from each region is aggregated. Subsequently, through four pyramid levels, the pooling kernels cover the entire image, half of the image, and progressively smaller regions. These pooled features are then fused with the global contextual representation and concatenated with the previous feature maps (Zhao et al., 2017).

DeepLabV3+ Architecture

Proposed by Chen et al. (2017), DeepLabV3+ represents a sophisticated advancement in the field of semantic segmentation. The architecture combines the Atrous Spatial Pyramid Pooling (ASPP) module with an encoder–decoder structure. The use of atrous (dilated) convolutions expands the model's receptive field without reducing spatial resolution or increasing the number of parameters.

Atrous convolutions extract image features while enlarging the receptive field, whereas the ASPP (Atrous Spatial Pyramid Pooling) module applies multiple dilation rates simultaneously, enabling the capture of contextual information at multiple scales. Feature responses are computed within Deep Convolutional Neural Networks (DCNNs) without requiring additional learnable parameters (He et al., 2023).

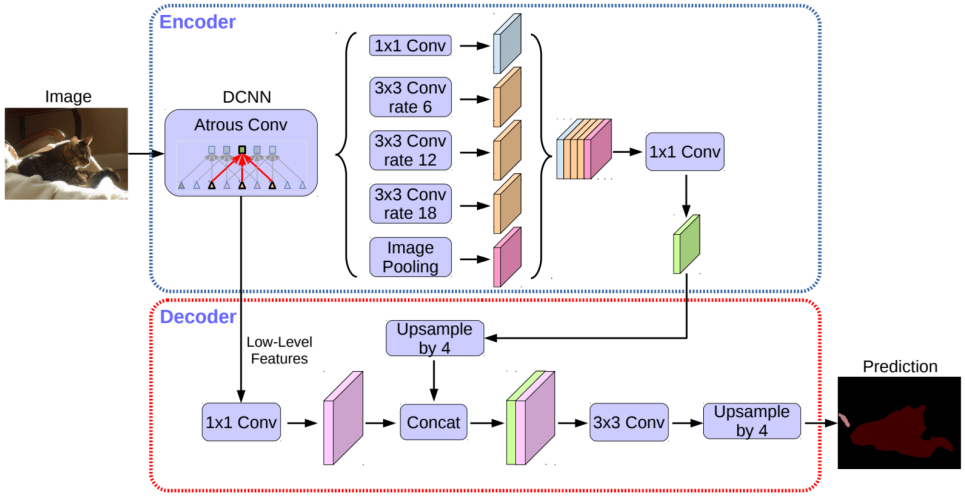


Figure 3. DeepLabV3+ Architecture Diagram

The backbone of DeepLabV3+ is MobileNetV2, a lightweight neural network specifically designed for deployment on mobile devices. It inherits the advantages of the MobileNet family while

replacing standard convolutions with depthwise separable convolutions. This process consists of two main stages: depthwise convolution and pointwise convolution.

In MobileNetV2, each input feature map is initially divided into multiple independent feature channels, and each channel is processed by 3 × 3 convolutional kernels, producing an equal number of output feature maps corresponding to those channels (Chen et al., 2017).

Materials and Methods

Image Dataset

For this research, a dataset consisting of 1,017 images of *Prata Catarina* banana plants was used, properly annotated for the segmentation of banana bunches and rachises. A sample of the dataset employed can be observed in Figure 5.

Image acquisition was performed using eight different smartphones, resulting in images with varying resolutions, namely: 3213 × 5712 pixels (*iPhone 15*), 3072 × 4080 pixels (*Redmi Note 11*), 3072 × 4080 pixels (*Redmi Note 12*), 3120 × 4160 pixels (*Samsung Galaxy A03s*), 2250 × 4000 pixels (*Samsung Galaxy A31*), 1884 × 4080 pixels (*Samsung Galaxy A54*), 6144 × 8160 pixels (*Moto G22*), and 2992 × 4000 pixels (*Moto G54*).

Figure 4. Sample images from the manually segmented dataset.

All images in the dataset were stored in the Joint Photographic Experts Group (JPEG) image format and were manually annotated using contour-based segmentation labels for the following classes: banana bunches and rachises (stalks). The segmented dataset, including images, labels, and metadata, is publicly available on the Roboflow platform.

Evaluation Metrics

The performance of the segmentation models was evaluated using metrics that quantify the quality of correspondence between the predicted segmentations and the reference annotations (ground truth). The metrics adopted in this study were Precision, Recall, Dice Coefficient, and Loss.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{3}$$

$$BCE = -\frac{1}{N}\sum[y.log(\hat{y}) + (1-y).log(1-\hat{y})] \tag{4}$$

$$Jaccard\ Loss = 1 - \frac{TP}{TP + FP + FN} \quad (5)$$

$$Loss = BCE + Jaccard\ Loss \quad (6)$$

Where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) represent the classification outcomes, and A and B correspond to the pixel sets of the predicted segmentation and the reference segmentation, respectively.

Precision measures the proportion of true positives relative to the total number of positive predictions, reflecting the model's ability to reduce false positives, which are associated with Type I errors. In turn, Recall represents the proportion of actual positives correctly identified, being sensitive to the occurrence of false negatives, which are related to Type II errors.

The Dice Coefficient evaluates the similarity between predicted and reference segmentations by quantifying the degree of overlap between them, with values ranging from 0 (indicating no overlap) to 1 (corresponding to a perfect segmentation).

Experiment Design

The three models were trained in a Linux environment using the Google Colab web platform. The implementation of the models and experimental procedures was carried out using the Python programming language (version 3.12.12), employing the Keras 3.10 and TensorFlow 2.19 frameworks, with support from the Segmentation Models 1.0.1 abstraction library as the foundation for the development and training of the neural networks.

The Google Colab environment configuration included an Intel Xeon CPU running at 2.20 GHz with 12 cores, 52 GB of available RAM, and an NVIDIA L4 GPU with 22.5 GB of memory.

For dataset storage and manual segmentation, the Roboflow platform was used. The dataset was partitioned such that 70% of the images were allocated for training, 10% for validation, and 20% for testing. Considering a total of 1,017 images, 713 were used for training, 102 for validation, and 202 for testing.

The models' hyperparameters were standardized to ensure a fair comparison among the architectures. The parameters are presented in Table 1.

Architectures	Num of Epochs	Batch Size	Input (in pixels)
Psp-Net	40	16	384 × 384
U-Net	40	16	384 × 384
DeepLabV3+	40	16	384 × 384

Table 1. Common hyperparameters used across the evaluated models.

A hybrid loss function widely used in image segmentation and binary classification tasks, known as BCE–Jaccard Loss, was employed. This function combines Binary Cross-Entropy (BCE) and Jaccard Loss. The combination operates at the pixel level, enhancing model performance by measuring the similarity between the model's predictions and the provided segmentation mask.

Results and Discussion

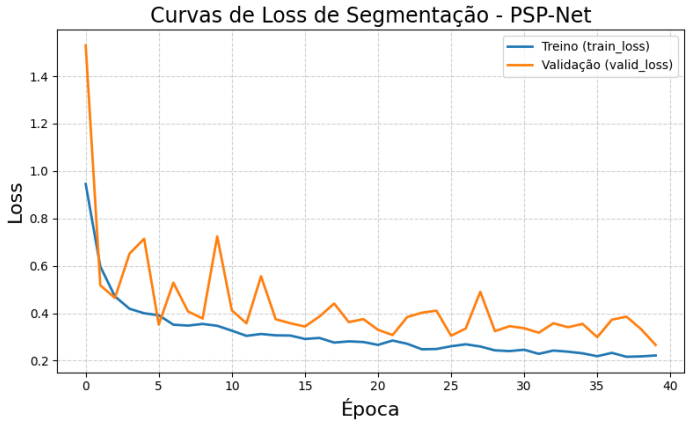
Training of the Models

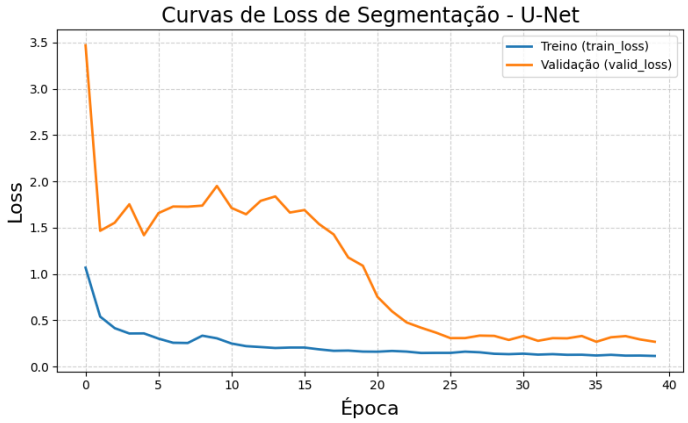
Table 2. Percentage results of the training metrics in the final epoch for each model.

Architectures	Train_Precision	Train_Recall	Train_Dice
Psp-Net	90,20%	89,55%	89,87%
U-Net	94,86%	95,31%	95,09%
DeepLabV3+	94,69%	95,14%	94,91%

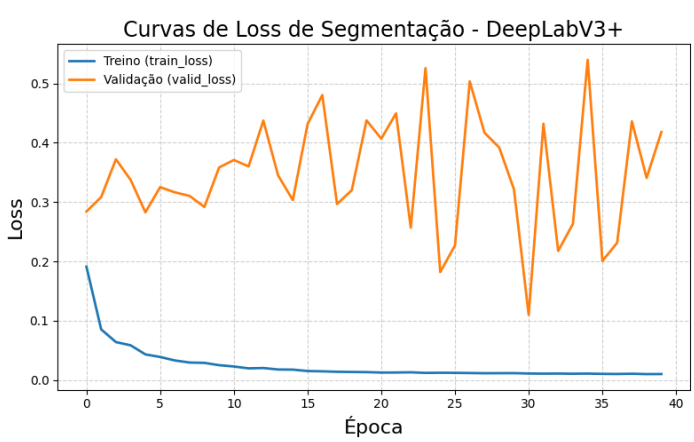
It can be observed that U-Net and DeepLabV3+ achieved very similar training results, with training-stage Dice scores close to 95%. This behavior indicates a high capacity for parameter fitting to the training data, suggesting strong adaptation and effective feature extraction from the provided samples.

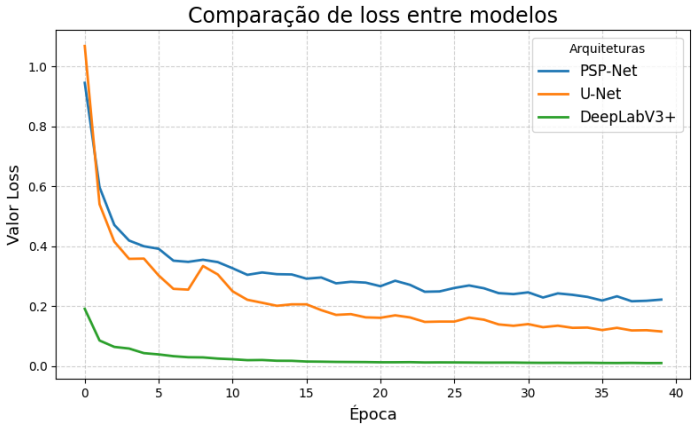
PSPNet, on the other hand, presented slightly lower metrics compared to the other models, exemplified by a Dice score of only 89.87%. Nevertheless, it demonstrated consistent convergence throughout the training epochs, as observed in the loss curves (Figure 5). This behavior may indicate that the model successfully captured the main characteristics of the banana bunch, although it still showed uncertainty regarding certain isolated features within the segmentation masks.





a) **b)**





c) **d)**

Figure 5. (a) Segmentation loss curves of the PSPNet model. (b) Segmentation loss curves of the U-Net model. (c) Segmentation loss curves of the DeepLabV3+ model. (d) Comparison of training loss curves for each architecture.

Validation of the Architectures

By analyzing items (a, b, c) in Figure 5, it is noticeable that during the first 20 epochs U-Net struggled in the validation stage due to insufficient information to distinguish the banana bunch from the surrounding environment. A similar behavior was observed in PSPNet, manifested as performance peaks and fluctuations. This instability may indicate that the models initially had difficulty generalizing the data, suggesting possible overfitting during the early stages of training.

In contrast, DeepLabV3+ exhibited pronounced instability in generalizing validation data, effectively learning only the training samples it was exposed to while failing to properly segment validation images. This behavior reflects severe instability and strongly indicates overfitting in the model.

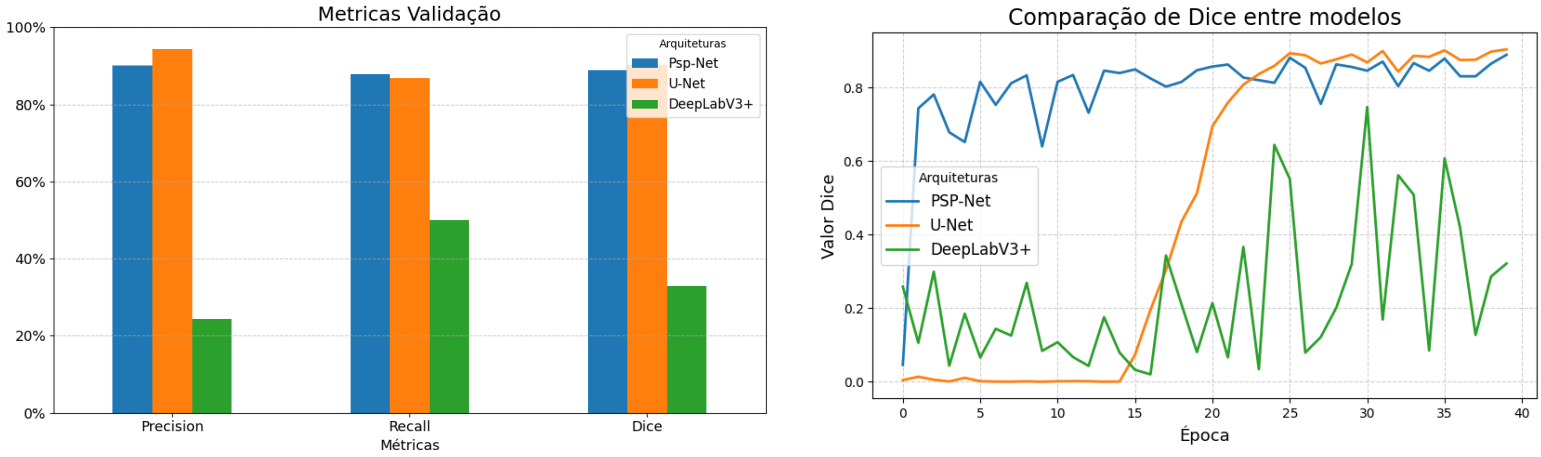


Figure 6. (a) Comparison of validation metrics in the final epoch for each architecture. (b) Comparison of the Dice score curve in the final epoch for each architecture.

As illustrated in Figure 6 (a, b), during the final validation epoch, U-Net demonstrated a considerable improvement in performance, standing out in terms of accuracy (94.37%) and achieving the highest Dice score (90.39%). This result is consistent with the literature, which highlights the effectiveness of this architecture in object segmentation tasks involving well-defined boundaries and moderately sized datasets, even when using parameters that are not fully optimized for object segmentation.

PSPNet showed balanced performance, achieving the highest recall value (87.77%), indicating greater sensitivity in identifying pixels belonging to the banana bunch. However, it presented a slight reduction in precision compared to U-Net, suggesting a higher inclusion of false positives.

In contrast, DeepLabV3+ exhibited significantly inferior validation performance, with a Dice score of only 32.83%, despite achieving training results similar to those of U-Net. This discrepancy suggests the possible occurrence of overfitting. The standardization of hyperparameters, required to ensure comparability among the models, may have negatively affected this specific architecture. Due to its use of atrous (dilated) convolutions and more complex Spatial Pyramid Pooling modules, DeepLabV3+ often requires a substantially larger volume of data and more refined hyperparameter adjustments—such as carefully tuned learning rate and batch size—to achieve proper convergence. Under the evaluated conditions, consisting of 40 training epochs and a limited dataset, the model likely failed to generalize effectively and learn the morphological characteristics of the *Prata Catarina* banana bunch, instead memorizing the manually segmented training data.

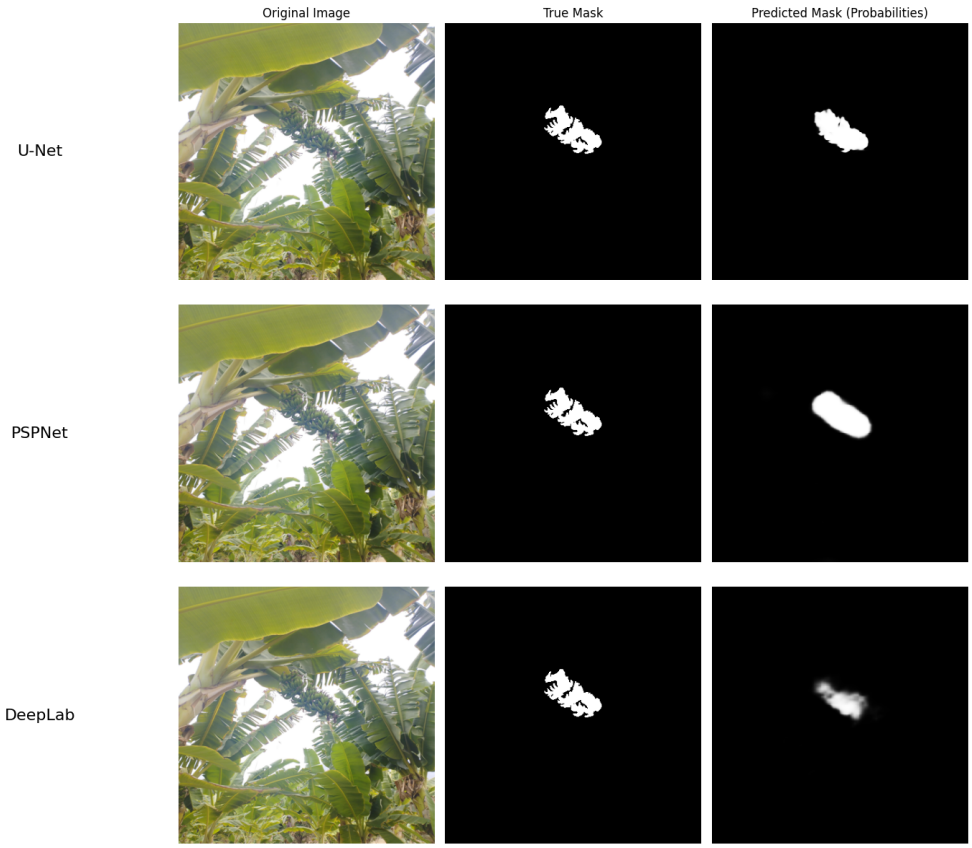


Figure 7. Comparative Prediction among the Three Models Using the Same Reference Image

The qualitative analysis of the segmentations shown in Figure 7 reinforces these findings: U-Net demonstrated greater accuracy in delineating the banana bunch; PSPNet achieved intermediate performance; and DeepLabV3+ exhibited evident difficulties, producing masks with a reduced segmented area and a significant presence of noise.

Conclusion

The present study demonstrated that the application of Deep Learning–based semantic segmentation techniques constitutes a viable and promising approach for automating bunch monitoring in *Prata Catarina* banana cultivation. Among the evaluated architectures, U-Net proved to be the most suitable for the proposed experimental scenario, achieving, in the final validation epoch, an accuracy of 94.37% and a Dice score of 90.39%. Its architecture, incorporating skip connections, showed a high capacity to preserve spatial information and more accurately delineate bunch contours, thereby improving the quality of the segmented masks.

PSPNet presented intermediate performance and consistent behavior, demonstrating good generalization capability, particularly in terms of recall, although with a slight increase in false positives. In contrast, DeepLabV3+, despite expressive training results, exhibited low generalization capacity on the validation set, suggesting the occurrence of overfitting.

Overall, the findings reinforce that the selection of an architecture should consider not only its structural sophistication but also the amount of available data, the complexity of the problem, and the computational resources employed. As future perspectives, the following are recommended: (i) expansion of the dataset with greater variability in lighting conditions and phenological stages; (ii) application of more robust data augmentation techniques; (iii) individualized hyperparameter tuning for each architecture; and (iv) testing in real embedded environments to enable field validation.

Therefore, it can be concluded that semantic segmentation applied to banana cultivation represents a strategic technological tool to advance precision agriculture, contributing to greater standardization, productivity, and competitiveness within the sector.

Acknowledgments

This study was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil – Finance Code 001. Danielo G. Gomes acknowledges the support of the National Council for Scientific and Technological Development (CNPq) under grant numbers 311845/2022-3 and 403996/2025-2.

References

Baglat, P.; Hayat, A.; Mendonça, F.; Gupta, A.; Mostafa, S.S.; Morgado-Dias, F. Non-Destructive Banana Ripeness Detection Using Shallow and Deep Learning: A Systematic Review. Sensors 2023, 23, 738. https://doi.org/10.3390/s23020738

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.05587

Elinisa, C. A., Wa Maina, C., Vodacek, A., & Mduma, N. (2025). Image Segmentation Deep Learning Model for Early Detection of Banana Diseases. Applied Artificial Intelligence, 39(1). https://doi.org/10.1080/08839514.2024.2440837

FAO Knowledge Repository. (2025). Fao.org. https://openknowledge.fao.org/handle/20.500.14283/cd2622en

F. F. Aradhana, M. D. Sulistiyo, E. Rachmawati, S. Hadiyoso, N. M. Z. Hashim and H. Murase, "Semantic Segmentation for Fruit Freshness Identification Using U-Net," 2024 12th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2024, pp. 160-165, doi: 10.1109/ICoICT61617.2024.10698186

G. García-Downing, A. Calderón-Barboza, J. L. Jiménez, L. A. C. Zamora and L. A. B. Artavia, "Deep learning system for detection and classification of banana and plantain cultivation zones in satellite images, implementing data parallelism," 2025 International Conference on Engineering Management of Communication and Technology (EMCTECH), Vienna, Austria, 2025, pp. 1-8, doi: 10.1109/EMCTECH65814.2025.11220582.

He, J., Duan, J., Yang, Z., Ou, J., Ou, X., Yu, S., Xie, M., Luo, Y., Wang, H., & Jiang, Q. (2023). Method for Segmentation of Banana Crown Based on Improved DeepLabv3+. *Agronomy*, *13*(7), 1838–1838. https://doi.org/10.3390/agronomy13071838

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe. *Proceedings of the ACM International Conference on Multimedia - MM '14*. https://doi.org/10.1145/2647868.2654889

L, B. A., OLIVEIRA, COSTA, ALMEIDA, J, A. E., F, C. E., MATSUURA, SOUZA, FANCELLI, M., MEISSNER, OLIVEIRA, SILVA, M, M. V., & CORDEIRO,. (2017). A cultura da banana. *Embrapa.br*. https://doi.org/85-7383-037-9

Lu, J.; Xiang, J.; Liu, T.; Gao, Z.; Liao, M. Sichuan Pepper Recognition in Complex Environments: A Comparison Study of Traditional Segmentation versus Deep Learning Methods. Agriculture 2022, 12, 1631. https://doi.org/10.3390/agriculture12101631

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1505.04597

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017, April 27). *Pyramid Scene Parsing Network*. ArXiv.org. https://doi.org/10.48550/arXiv.1612.01105